Exploiting Social Networks for Recommendation in Online Image Sharing Systems

Thesis

THE OPEN UNIVERSITY

Exploiting Social Networks for Recommendation in Online Image Sharing Systems

by

Adam Rae

A thesis submitted for the
degree of Doctor of Philosophy

August 2011

Contents

List of Figures

List of Tables

List of Publications and Patents

Papers

Improving Tag Recommendation Using Social Networks - RIAO 2010

Adam Rae, Börkur Sigurbjörnsson and Roelof van Zwol

This paper resulted from the experimentation and results found in Chapter 3 on mining tag networks derived from social interactions. I show the relative ability of tags from different subsets of photos (based on social and non-social criteria) when making suggestions for users while they are tagging in online image sharing systems like Flickr. I led the execution of the work presented in this paper and undertook the analysis of our findings with my co-authors.

Prediction of Favourite Photos using Social, Visual and Textual Signals - ACM MM 2010

Roelof van Zwol, Adam Rae and Lluís Garcia Pueyo

Based on the experimentation described in Chapter 4 on learning user preference using multiple features types from the Favourite label in Flickr. I showed how features based on social interaction between users had a significant impact on the performance of my machine-learned approach to favourite image categorisation. I led the experimental design and execution of work in this paper and undertook the analysis of findings with my co-authors.

Exploiting Term Co-occurrence for Enhancing Automated Image Annotation - Evaluating Systems for Multilingual and Multimodal Information Access 2008

Ainhoa Llorente, Simon Overell, Haiming Liu, Rui Hu, Adam Rae, Jianhan Zhu, Dawei Song and Stefan Rüger

As a result of our participation in the Visual Concept Detection Task at ImageCLEF 2008 in which I contributed to the execution of the experiments, this paper describes the advantages of using semantic ontologies to augment probabilistic image analysis. It also introduces the statistical co-occurrence technique that I extend in Chapter 3.

Geographic and Textual Data Fusion in Forostar - Evaluating Systems for Multilingual and Multimodal Information Access 2008

Simon Overell, Adam Rae and Stefan Rüger

This work describes techniques for fusing textual and geographic relevance information in geographic information retrieval using a robust filter-based approach that outperformed a more sophisticated penalisation method. I built and helped evaluate the rank combination mechanisms used in Simon Overell's Forostar system, experience of which led to the techniques I used in Chapter 3 to combine tag suggestions from complementary tag graphs.

Reviewing multimedia search engines - "Inside Knowledge Magazine"

Stefan Rüger and Adam Rae

An article for a non-academic industry magazine that gave an overview of the state of the art in multimedia information retrieval (MIR) in which I contributed (amongst others) a section on the value of understanding the social context of a user when fulfilling their information needs, a position reinforced by my findings in both of my experimental chapters.

Patents

I am a co-inventor on the following published patent based on the work of Chapter 3:

> "System for Personalized Term Expansion and Recommendation", *Application number: 12/537,157, Publication number: US 2011/0035350 A1, Filing date: August 6, 2009*

I am also a co-inventor on the following pending patent based on the Flickr Favourites techniques described in Chapter 4:

> "Predicting Photo Favourites", *Submission date: July 22, 2009*

Acknowledgements

This thesis started after soul searching about which direction I wanted my life to take. I realised I had been on a trajectory throughout my education, aiming for this particular milestone and that it felt natural to pursue what would (I hoped) be a rewarding experience.

If I had not taken and so enjoyed Stefan Rüger's course on multimedia algorithms and techniques as an undergraduate, I probably would not have plucked up the courage to apply for a studentship with him. As my supervisor, he has shown encouragement, guidance and support throughout this endeavour and for this I thank him.

While our time working together was relatively brief, I would also like to thank Dawei Song for his support in the early months of my PhD.

Roelof van Zwol, also a supervisor, reinforced my excitement for the field we work in—his energy and vision have been vital to this work, for which I am immensely grateful.

I was lucky that Suzanne Little was able and willing to join my supervision team part way through my studies and I value not only her professional input to my work, but also her friendship as someone I have had the privilege to work with closely. We will build that super-amazing-DIY-touch-table one day!

I would also like to thank Marian Petre and Cathal Gurrin for agreeing to scrutinise and examine my work.

Funding for this work come from an EPSRC CASE collaboration with Yahoo! Research and Fundació Barcelona Media to whom I am grateful for their support.

While based in Barcelona at Yahoo! Research I was fortunate to have the opportunity to work along side Börkur Sigurbjörnsson and Lluís Garcia Pueyo. I will continue to be envious of Börkur's razor-sharp insight into problem solving and Lluís' well-honed skills. I also count my self lucky to have have had them as guides to Barcelona as well as friends.

Thanks must also be given to Joe Corneli who gave enthusiastic advice and support when I was formulating the mathematical description of the model in Chapter 3.

I thank Ricardo Baeza-Yates for enabling such fruitful collaboration between The Knowledge Media Institute (KMi) and Yahoo! Research Barcelona and for his understanding and patience near the completion of this thesis. On this note I must also thank my friend and colleague Vanessa Murdock, who not only gave valuable guidance throughout my initial time in Barcelona, but based on that experience went on to let me work with her in a postdoctoral-position, being patient and supportive during the difficult overlap period between my studies and work.

Life as a PhD student is not and really should not be lived in isolation and I am immensely grateful to the other students in KMi without whom battling on through such a task would have been many times more difficult. I'd like to thank Haiming, Sofia, Fouad, Miriam, Carlos, Rui and so many more, you're all very special people. In particular, I must thank Ainhoa for being such a good friend, who has always had time for a cup tea and a chat, on occasion the most valuable gift someone can give.

I must also thank the wider postgraduate community at the OU. We are a small group, but I think the maxim regarding quality and quantity is well suited to describe you all. I shall miss you and the evenings in the cellar bar.

My final thanks must go to my family, who have been immensely supportive throughout the PhD process. Just being there during the times when things seemed insurmountably tough as well as for each minor achievement is only part of why I'm so grateful. Dad, Emma, Peter, Olga, Richard and Anna, this thesis would not have been written without you.

This thesis is dedicated to my mother, who always saw my potential and did all she could to help me see it too.

Rwy'n cyflwyno'r thesis i fy mam.

Abstract

This thesis aims to demonstrate the distinct and so far little explored value of knowledge derived from social interaction data within large web-scale image sharing systems like Flickr, Picasa Web, Facebook and others for image recommendation. I have shown how such systems can be significantly improved through personalisation that takes into account the social context of users by modelling their interactions by mining data, building and evaluating systems that incorporate this information. These improvements allow users to search and browse large online image collections more quickly and to find results that more accurately match their personal information needs when compared to existing methods.

Traditional information retrieval and recommendation datasets are contrived to provide stable baselines for researchers to compare against but they rarely accurately reflect the media systems users tend to encounter online. The online photo sharing site Flickr provides rich and varied data that can be used by researchers to analyse and understand users' interactions with images and with each other. I analyse such data by modelling the connections between users as multigraphs and exploiting the resultant topologies to produce features that can be used to train recommender systems based on machine learnt classifiers.

The core contributions of this work include insight into the nature of very large-scale online photo collections and the communities that form around them, as well as the dynamic nature of the interactions users have with their media. I do this through the rigorous evaluation of both a probabilistic tag recommendation system and a machine learnt classifier trained to mimic user decisions regarding image preference. These implementations focus on treating the user as both a unique individual and as a member of potentially many explicit and implicit communities. I also explore the validity of the Flickr 'Favourite' feedback label as proxy for user preference, which is particularly important when considering other analogous media systems to which my findings transfer. My conclusions highlight how vital both

social context information and the understanding of user behaviour are for online image sharing systems.

In the field of information retrieval the diverse nature of users is often forgotten in the hunt for increases in esoteric performance metrics. This thesis places them back at the centre of the problem of multimedia information retrieval and shows how their variety and uniqueness are valuable traits that can be exploited to augment and improve the experience of browsing and searching shared online image collections.

Declaration of Authorship

I, Adam Rae, declare that this thesis titled, 'Exploiting Social Networks for Recommendation in Online Image Sharing Systems' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at The Open University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Note Regarding URLs

All URLs (Uniform Resource Locators) used in this thesis were checked and confirmed to be correct at the time of printing. However, online resources may change over time, and where possible, more persistent resources are referred to.

Chapter 1

Introduction

> *"We are now making history,and the sun picture supplies the means of passing down a record of what we are, and what we have achieved in this nineteenth century of our progress..."*
>
> (John Thomson, 1891, official photographer to King George V)

It has become a cliché to start a discussion on the field on online multimedia information retrieval by stating just how meteoric the growth of data has been in recent years. While it may be an overused way of introducing work in this field it is, nonetheless, true.

For example, in September 2010, the online photo sharing website Flickr reached the milestone of 5 billion uploaded images (Sheppard, 2010), while Facebook reached 10 billions items[1] a couple of years earlier. To put that in context, the British Library has a total of just over 150 million items (Office, 2010) and the American Library of Congress has over 147 million items[2], covering books, manuscripts, maps and sound recordings.

With the proliferation of ever cheaper digital cameras, and their diversification away from single-use devices like cameras and into everything from computer webcams to mobile phones and tablets, it has become easier than ever before to take digital images. The Internet has

[1] http://www.facebook.com/note.php?note_id=30695603919

[2] http://www.loc.gov/about/generalinfo.html#2010_at_a_glance

also made it easier for photographers to share their pictures with friends and family as well as complete strangers via the online world.

Systems designed to handle effectively the billions of photos and images Internet users upload online each year continue to proliferate. New ones are created based on new technology, gimmicks and brands, with many withering away while a few go on to prosper. And with the rise of the 'socialisation' of web activity, those media sharing sites that thrive have become combinations of both technology and community.

So although social activity has become associated with online media sharing, the value of the social interaction information afforded by such integrated systems has not yet been fully explored nor effectively exploited. Understanding the motivations and needs of the people who use such systems, how and why they interact with others and what this tells us about what they want is difficult set of questions. It is only by addressing them that better, more effective systems can be built to satisfy users who share their photos online.

To help break this problem down, a fundamental distinction must be drawn between two common types of interaction users have with online photo sharing systems, namely *searching* and *browsing*. The first is well understood d(yet still very difficult) problem of taking a request for information from a user, interpreting it and retrieving the media in the system that most closely satisfies it. The second is the less well defined task of discovering content, not necessarily with a specific information need in mind, other than perhaps to find something that piques the browser's interest. What that interest may be can be influenced by many factors, ranging from general topical interests of the user to their mood.

Moreover, photos are unlike the text documents that the field of information retrieval has been so focused on in the past. They are visual media that can have a far more immediate emotive impact—reportage from war zones to striking art photography can make an impression on viewers very quickly. Perhaps most of all we find ourselves drawn to images that have a personal aspect, such as those of our friends and family.

In the past photographers may have kept their photos in the equivalent of a shoebox under the bed and, for those more conscientious among them, with carefully written notes on the back of exactly which relatives are in the photo and at which big family event it was taken at, etc. The photos were difficult to share widely and their annotations were dependent on the photographer's knowledge and memory.

Nowadays users can upload images to the Internet direct from the capture device (in the case of mobile phones), encode pertinent contextual information automatically and, perhaps most importantly for them, make it available to others and allow them to annotate and generally interact with the image as well. Digital photography has transitioned from being a solitary endeavour to an interactive group activity.

This social aspect of image usage is growing more prevalent and yet the systems we use to handle them online are only beginning to truly take advantage of this newly available contextualising data.

This thesis focuses on the activity of recommendation in image-sharing websites (with a particular focus on Flickr) and what role social context information plays in both enhancing current tag suggestion systems (Chapter 3) and extending and enriching personalised image recommendation (Chapter 4), which I interpret as task for predicting favourite images.

While the theory of recommender systems has been explored before, to the author's knowledge this is the first time multiple personalised social graphs have been effectively combined for use in tag suggestion, as well as the first time social context features have been integrated into a multi-modal approach to image recommendation.

1.1 Motivation

Multimedia information retrieval, and its sub-field of image retrieval, focuses on the effective indexing, storing, retrieval and presentation of media to users. Recommender systems research focuses on the retrieval and presentation aspects in particular, even though it can

have an influence on the other two areas (predictive caching strategies and automatic annotation for example).

A key aspect of the area of retrieval is the matching of the user's information need to the data available, a process that requires interpretation by the system. In straight-forward search use-cases with well structured and annotated data, strictly defined query languages and effective matching mechanisms, information retrieval systems can work well and satisfy users to a high degree.

However, there are also less well-defined use-cases in which users expect a wider range of interaction, more than just the translation of a textual query into a result set. These use-cases include content discovery where users wish to find media that interest them and that they have not seen before, and browsing, where users navigate a media database (or more usually subsets thereof) in a continual process in which their information need changes and evolves as they are exposed to more media. In both of these cases, the system has to tailor results to specific users and involves a different kind of interpretation than plain query-based search, and it is far more difficult to judge success.

Contextual information regarding users, their attributes, their typical behaviour and their social interactions is useful in guiding users through data to get them to what will satisfy them. Unfortunately, users rarely provide explicit and comprehensive feedback to the system allowing it judge how successful it has been, making it difficult for the system and for researchers to evaluate its effectiveness. Even when prompted for direct feedback on their satisfaction with respect to the relevance or interest of the data shown to them, it is difficult to gather objective data for what is, to the user, a subjective task.

Techniques for effectively increasing user satisfaction when browsing and discovering content ultimately benefit the browsing users themselves, but can have significant knock-on effects for other use-cases. For example, by improving the quantity of high-quality of metadata associated with images, and hence making it easier for the system to accurately cluster images with similar semantic content for recommendation while browsing, users who issue

search queries benefit as well. By predicting which images that freshly enter the database particular users are likely to find interesting, search interfaces can extend the presentation of results specific to a search query with those that may also be valuable to the searching user.

There are a number of identifiable groups of beneficiaries of more effective recommender systems for large-scale image collection. These include, but are not limited to:

Flickr users As the experiments undertaken in this research were done using data gathered from Flickr, my findings are most immediately relevant to the user of that specific site. As one of the web's largest photo sharing systems, improvements to its systems based on my work would have the possibility of having an impact on their over 30 million users around the world.

Other photo-sharing site users While Flickr is indeed a one of the most popular photo-sharing sites, there are others, including Facebook and Google Picasa Web Albums. Particularly in the case of the former, which specialises in social interaction among its 800+ million users[3], and for which photos and other media are a secondary focus, the greater richness of social data available in that system could potentially lead to even more significant improvements in users' satisfaction than with Flickr.

Other media-sharing site users Photos are but one form of multimedia data generated and shared online. Video and audio are also being shared more than ever before and sites like YouTube and Vimeo could adopt the social graphs and social features I used in this experimental work to improve the annotation of their media as well as improve media recommendation by taking into the social context of their users.

[3]Number as of August 2011, source: http://www.facebook.com/press/info.php?statistics/

Non-internet based media collection users More speculatively, any system where users need to browse through media collections, in museums, libraries, etc., may be able to take advantage of socially-aware recommender systems. As an example, a museum's digital library could perhaps help guide students to materials they will find interesting and useful if it is also aware of what their classmates have found valuable. Or a DVD rental service would benefit greatly by using information about the social connection between its subscribers in order to make targeted suggestions that are more likely to lead to more custom.

1.1.1 The growth of online media sharing

Personal photo collections have moved away from physical prints that record scenes using chemical compounds and towards digital files encoding light in bits and bytes. This change has altered the way users (or photographers, image creators, etc.) are able to use their images, now being able to copy, manipulate and transfer them in new ways. This has meant that users have been able to share images—no longer physically, slowly and expensively, but digitally, more quickly and for almost negligible cost. This has allowed users to share their media via narrowcasting mechanisms like email as well as broadcasting them via personal websites. Centralised systems have developed that allow many users to have a single place to upload and share their media with each other or specific sub groups.

This thesis is concerned predominantly with Yahoo's Flickr photo sharing website. As one of the largest photo sharing websites on the web (at the time of writing) that also incorporates social activity, it is an excellent source of data for the experimental work undertaken in Chapters 3 and 4.

1.1.2 The importance of recommender systems

Connecting users with the photos in online digital collections that they seek or will interest them is a non-trivial task. Being able to guide a user in their task by relevant, targeted

suggestions is an excellent way to shorten the process and increase the likelihood of user satisfaction. These suggestions are most famously already in use in systems other than online photo-sharing systems, such as supermarket direct market coupons that give discounts on products customers are likely to want based on their past shopping habits and product suggestions in online shops like Amazon.com based on what other people who exhibit similar interests have also found relevant.

1.1.3 The weaknesses of current approaches

While existing techniques for recommendation in photo sharing websites do exist, they tend to use information aggregated over the whole collectivity of users who interact with the system. This is used to make recommendations based not on the needs of the individual user but on the community as a whole, with the expectation that, for most cases, this will be sufficient. And, in many cases, it is, but this approach ignores the variations in interests and behaviour between users and fails to satisfy all users.

This naturally leads to the topic of personalised recommendation, which makes recommendations for specific users (or classes of users). The advantages and disadvantages of this form of personalisation are discussed in Chapter 2, but, to summarise, these approaches benefit from as much contextualising information about a user as possible. At present existing recommendations systems focus on characterising the media itself (in terms of textual and content-based features) and the attributes of user (such as demographics). They do not currently, however, take into account the social context of the user—their implicit and explicit connections to other users and how they interact with them.

1.1.4 The potential of social context awareness

As users become more easily able to connect and interact with others online, and as users want to do this more, the data this generates is also increasing. The profile of a user with

respect to the usage of a photo sharing website is no longer made up of information that describes them in isolation like their attributes and personal activity, but also in terms of their explicit and implicit connections and interactions with other users.

While photo retrieval and browsing are inherently difficult tasks, recommender systems attempt to ameliorate these problems. However, there are currently no optimal solutions that consistently satisfy all users. In addition, existing solutions do not exploit the new wave of social context data now available. Can this data, therefore, be used to improve existing recommendations systems used in online photo sharing, such as tag and image suggestion?

1.2 Hypothesis

This thesis is based on an hypothesis which I test using a series of empirical experiments. These experiments are designed around a number of questions that I answer partly through my exploration of related work in Chapter 2, but mostly through the work on tag suggestion using social graphs in Chapter 3 and on multi-modal Flickr image recommendation incorporating social context data in Chapter 4.

1.2.1 Main hypothesis

Existing techniques for managing large-scale online image collections do not currently fully take advantage of the rich social context of the data itself and the users who interact with it, a form of data that is increasingly available. Nor do they leverage the social connections between people who use such systems. By accurately modelling these connections and understanding more about them, we learn more about user image and tag preference. *More specifically, image and related metadata recommender systems can be built using this social information that are more effective than existing state-of-the-art non-social techniques.*

I use the term 'effective' here to mean a system that can make recommendations that users within the diverse community found in systems like Flickr agree (or a suitably analogous

proxy evaluation task suggests) that it is a suitable recommendation. This must be considered within the context of the task at hand.

1.2.2 Sub-questions and breakdown

The following questions are addressed throughout this thesis so that when they have been sufficiently answered individually, they will have provided evidence to test my main hypothesis.

1. **Which social connections yield the most valuable information for use in tag and image recommender systems designed for large online photo sharing systems?** Users form many explicit and implicit connections both directly and indirectly, depending on the interactions mechanisms available to them in photo sharing systems. Which of these links are most highly discriminating when used to train a recommender system for tags or images in such an environment?

2. **How can these social connections be effectively used to improve recommendation?** Mining data can lead to insight into trends and patterns in user interaction behaviour, but this has rarely been applied to the specific task of recommendation in image collections. I suggest that machine learning techniques can be used to learn such trends and form the basis for social context aware recommender systems for photos.

3. **How can different kinds (textual/visual/social) of media/user descriptors be combined effectively in an image recommender system?** Different aspects of media (their content, their semantic metadata, the information about the owner, etc.) provide different cues as to their relevance for particular users. But it is not currently known when these individual aspects are most useful, not how much this can vary between individual users. I propose that a suitable balance of these aspects can be automatically derived and furthermore, it can it be derived in advance of the moment

of recommendation. This can be shown by using the task of relevance prediction and in so doing derive techniques that maximise performance for the kind of diverse community found on sites like Flickr.

4. **Can single positive feedback cues like the Flickr 'Favourite' label be used to train systems to predict further 'Favourites'?** In many online media sharing environments users are able to annotate images with a tag or label of approval—'Favourite' in Flickr, 'Like' in Facebook, Youtube and '+1' for other Google products, etc. These are sometimes the only method of user preference feedback available, meaning there is no negative feedback available and no grading of preference expressing the extent of approval. Can user image preference as inferred by such feedback be accurately modelled using only this single, positive, cue in the case of Flickr?

Definition: The use of the word *favourite* to describe an image will be used throughout this thesis, particularly in Chapter 4. This word will be used in two ways.

The first refers to the general definition of an image or set of images that are preferred over others by a particular user and will be used in non-capitalised form. The second refers to the specific use of the binary label "Favourite" used in the Flickr online image sharing system with which users can annotate any image in the system they view. This will be used in the capitalised form.

It should also be noted that as Flickr is a predominantly American website they favour the US spelling "Favorite", however for consistency with the orthography of this thesis, the spelling "Favourite" will be used here.

1.3 Contributions

1.3.1 Suggesting tags for photos using personalised data and social graphs

This set of experiments extends existing collective techniques for suggesting tags to users as they are annotating their photos with personalised data from social interactions. I model the connections between users as graphs, specifically for the 'Contact' relationship (an explicit label between users) and the 'Group' relationship (where users are members of the same Flickr group). I then model the co-occurrences of tags that are used to annotate photos of individual users, their two social graphs as well as the community as a whole.

I then use this co-occurrences information with a probabilistic model to make suggestions for relevant tags on a per user basis. I show in which cases my combined approach supersedes the performance of existing, non-social techniques with respect to Mean Reciprocal Rank, Mean Average Precision and Precision at 5. With my co-researchers, we analyse the results to show the relative contribution of each personalised tag co-occurrence graph and break down the results with respect to the activity attributes of Flickr users.

Chapter 3 demonstrates the value of social data for the specific task of tag suggestion, as well as highlighting the importance of knowing exactly which aspects of this data can be useful for similar tasks. Understanding this data in finer detail is explored in the next chapter.

1.3.2 Predicting Flickr Favourites using social context information

Using the findings of the previous set of experiments, Chapter 4 proposes and implements a multi-modal approach to tackling a related recommendation problem, that of image recommendation. In the absence of explicitly labelled non-Favourite images, I propose and use two datasets that attempt to fulfil that role with respect to training.

I show how multiple features extracted from images, from their uploaders and from their social context can be used by a machine learnt mechanism to predict which photos in an incoming stream of previously unseen images a particular user is likely to label as a Favourite.

I analyse the relative value of content-based, textual and social image features and show that for those features being analysed, and for the negative data training set which more accurately reflects real Flickr usage, the social ones tend to be the most valuable, varying in accordance with a user's social activity. I then show how and in which cases individually trained machine learnt classifiers can be more highly performing than general, one-size-fits-all classifiers.

1.4 Thesis organisation

This introduction to my thesis describes the problem of media recommendation and the potential value of social context information when applied to this problem. It also outlines the hypothesis that underpins the research and experimental work undertaken in later chapters. The other chapters address specific topics within this research as outlined here.

Chapter 2 - Social context use in digital image recommendation systems This chapter explores the fields of social media, recommendation and comprehensively reviews the current state-of-the-art regarding features commonly extracted from media and specifically photos online for information retrieval purposes. The discourse is focussed on those areas that are particular relevant to the experimental work undertaken in later experimental chapters.

Chapter 3 - Personalised Tag Suggestion Using Social Context Exploiting social context using the established personalisation mechanism of media metadata (tag) suggestion is introduced in this chapter. I model user interactions with respect to their photos in online media sharing platform Flickr, as a number of well defined social graphs. I then show how

these graphs can be efficiently traversed to provide data that can be used to train a tag recommendation system that outperforms existing state-of-the-art techniques.

Chapter 4 - Identifying Flickr Favourites Using Social Context Based on the findings of Chapter 3, I extend the catalogue of social features that characterise users, their media and their interactions with their media. I then apply these to a second social recommendation scenario whereby I predict Flickr images for users that they are likely to label as Favourites. I evaluate and analyse the system presented and show how social features are valuable both by themselves and as a useful addition to features used in existing systems.

Chapter 5 - Conclusions I summarise the findings of this thesis and evaluate its success with respect to my hypothesis. I discuss the limitations of my work and in so doing propose a number of research questions that may be suited to further investigation.

Chapter 2

Social Context Use in Digital Image Recommendation Systems

> *"The trouble with having an open mind, of course, is that people will insist on coming along and trying to put things in it."*
>
> (Terry Pratchett, Diggers)

Roadmap This chapter explores the existing research in the area of social data use in media recommendation and demonstrates the place of my research with respect to the state of the art.

Section 2.1 looks at the history of digital imaging and how such data has developed from a isolated activity between solitary users, to a global and social endeavour.

Section 2.2 then looks at the field of recommender systems and shows the main classes of mechanism commonly employed, as well as introducing a new approach that I show in later experimental chapters to outperform existing techniques by using social context data.

This kind of data is explored in more depth in Section 2.3 where different forms of social data are outlined and their value to recommendation systems is emphasised.

As part of the multi-modal approach I adopt in experiment in Chapter 4, research on text based features derived from media shared online is looked at in Section 2.4 and image content features in Section 2.5.

Section 2.6 then looks at the role media datasets play in this field of research. I propose criteria against which such sets should be judged and catalogue a number of datasets used for media recommendation.

2.1 Photography from paper to pixels

While the chemical process of recording photographic images has been possible in one form or another for almost two hundred years, it took until the early 1960s for technology to develop to the point where researchers considered using electronic means.

As with many of the world's most profoundly important technology of the 20th century, the concept for the core component of modern digital cameras was born out of the needs of the space race. Eugene F. Lally of the American Jet Propulsion Laboratory in 1961 proposed a system that would allow the electronic recording of light using a an array (or 'mosaic') of photoreceptors that could convert the incoming analogue signal into a discrete, digitised representation (Lally, 1961). Such a system would allow a spacecraft to observe its surrounding starscape and navigate its way to Mars. Unfortunately, the electronics did not exist until the mid 1970s to produce such a device.

The Charge-Coupled Device, the core of most digital camera equipment used today, was first created as a form of computer memory, but its sensitivity to alpha particles made it unsuitable. However, this sensitivity to radiation made it an excellent imaging device.

By 1974, the company Fairchild Electronics had produced the first commercial charge-coupled device (CCD) suitable for imaging, with an array of 100 x 100 pixels (Williamson, 2009). By 1975, Steven Sasson and his team at the Eastman Kodak Company assembled the first digital camera using a CCD and demonstrated its viability. This device heralded the beginning

of digital still imaging, with the technology continually getting smaller, allowing far higher resolutions, at high levels of noise-free sensitivity. The digital signal processing to support such devices has also developed to make them flexible and robust. These electronic imaging devices have been used in applications ranging from astronomy, microbiology and robotics, to art photography and taking snaps for the family photo album.

As computers become cheaper and more powerful, the storage and processing of digital images become more feasible. Early computerised axial tomography (CAT) scanning produced digital images of slices of three dimensional object, providing a new, powerful tool useful in diagnostic medicine, an analytical tool in archaeology, a forensic tool for law enforcement and many others (González and Woods, 2008).

While these early systems indexed relatively low volumes of images, the need for systems that could scale better and match the needs of larger organisations became pressing. Examples of early large-scale image-handling systems include work by IBM on bank cheque scanning, processing and archiving in the system described in their 1994 patent (Yeskel, 2000).

Image capture technology continued to shrink and became cheaper to the point where digital imaging become affordable in the home as well as in the lab. People who were familiar with traditional chemical photography were now able to transfer their skills to the digital domain. Early devices produced low resolutions, for example Apple's QuickTake from 1994 (see Figure 2.1) could take images of 0.3 megapixels or million pixels (compared with the Nikon D3x announced December 2008 which can take images of 24.5 megapixels).

The advantage over film-based cameras however, of being able to take photos at essentially zero cost, caught on, and although the first commercial digital cameras had shortcomings, the early adopters were willing to overlook them, and their popularity grew.

FIGURE 2.1: Example of Apple QuickTake 100 digital camera, an early digital camera available to home users from 1994. From Flickr user *Jaqian_*, CC Attribution-ShareAlike 3.0

With the development of CD writing drives for home computers and colour inkjet printers capable of reproducing the digital images, digital cameras' presence in the home became commonplace. Whereas previously digital photography was a specialist and expensive hobby, the technology enabled its spread so that it became a common activity for home consumers and a cost-effective tool for business and commerce. This democratisation of digital photography opened new, wider markets for digital camera makers, who exploited the desire of users to be able to quickly and cheaply make photographs. At this time, the sharing of these images was predominantly through hard copies that had been printed out, either at home or at traditional photo development business who had printing capabilities. These were then disseminated in the same way film-based photos had been previously.

As internet connections become more common in the home, people started to share their digital photos with others. This was via methods like email, message boards and web-based home pages. The desire to share was facilitated by the low cost of transfer and the ease with which many people, regardless of location, could view images, particularly when compared to traditional film-based photography.

Flickr[1], the by-product of a massively-multiplayer online role-playing game (MMORPG), was developed by a company called Ludicorp in February 2004 (Graham, 2006) and provided an environment for users to upload, store and expose their photos online.

[1] http://www.flickr.com/

FIGURE 2.2: A number of websites provide photo sharing facilities, with varying services and functionalities. Logos are the property of their respective owners.

What started as a simple tool for collecting existing photos already on the web, became a platform for users to upload and share their own, and it is this focus that the site maintains today. Since its inception, Flickr has been joined by other similar sites (see Figure 2.2).

Existing systems characterisation and evaluation

I define a large-scale online photo-sharing system as one that satisfies the following criteria:

- Allows users to upload images from their own digital devices to the web service
- Once uploaded users can share their picture with other users, either completely or selectively
- Such a system has in the range of millions of actively participating users
- These users are able to interact with each other, either explicitly or implicitly

Systems like Flickr (owned by Yahoo!), Facebook and Photobucket are among those that satisfy these criteria. Table 2.1 shows an overview of these systems and their scale, with data from published sources (Facebook, 2008, 2010; Champ, 2009; Flickr, 2009; Photobucket, 2010).

	Flickr	**Facebook**	**Photobucket**
Number of user accounts	32m	500m	99m
Number of unique images	4b	10b	7b[2]
Date started	February 2004	February 2004	2003
Overview	Predominantly for sharing images, users can interact via comments, tags, groups and fora	A personal profile directory that allows users to share images, focussed on social connections	Photo upload and hosting service, particularly popular for hosting Twitter images

TABLE 2.1: Statistics of some of the larger image sharing websites in millions (m) and billions (b). All data taken from 2007-2010

With such large numbers of images being uploaded to these systems, and with so many people having access to them, the problem of effective storage, browsing and searching becomes even more important.

The approach taken by many early large-scale image-handling systems was to directly borrow techniques from the field of physical document information retrieval and adapt them to the specific needs of image indexing. This involved treating images solely as artefacts of their metadata: their manually created textual descriptions, tags and keywords and their device-generated EXIF data that encoded the state of the camera at the time of capture.

With relatively small, manually curated collections like those found in some digitised photography archives like the Fratelli Alinari Archive[3] and within photography using organisations like local newspapers, this metadata could be generated at a rate that kept up with new additions. Larger organisations like national archives, including the British Library[4] and the United States Library of Congress[5], as well as news agencies like Reuters and the BBC, have larger, manually annotated collections. For example, the Getty Images[6] has over 80 million photos, some originally digital, others that have been digitised, all of which have manually edited metadata.

[2] Figure for 'digital assets' which may include other media.
[3] http://www.alinari.com/
[4] http://www.bl.uk/catalogues/photographs/
[5] http://www.loc.gov/pictures/
[6] http://www.gettyimages.com/

However, the time it takes for a professional image librarian to create an image metadata description is not trivially short, and there are limits to how far this approach can be taken in terms of the cost involved in paying people to do this work.

In community image-sharing sites like Flickr, with billions instead of millions of images to handle and more being added every minute, this problem was addressed partly by relying on the users who submit images to annotate them themselves. This meant that professional annotators were no longer required. However, it also had the side-effect of reducing the overall quality of metadata (in terms of its ability to effectively discriminate images) for researchers who were re-purposing the data for other uses. This was due primarily to users not having the same objectives as researchers when it came to tagging their media. They spent little time comprehensively tagging/describing their photos and this made their annotations incomplete and inconsistent, compared to professional annotations.

2.2 Recommender and suggestion systems

2.2.1 Motivation

> "We are leaving the age of information and entering the age of recommendation." *The Long Tail, Anderson (2008)*

Recommender systems exist to address the problem of selecting a subset of information items from a larger collections such that the subset is more useful, relevant or of interest to a particular user. In their book *Recommender Systems: An Introduction*, Jannach et al. (2010) describe the fundamental function of all recommender systems as systems that support decision making. In non-digital contexts this kind of support task has traditionally been performed by trained people, from the librarian asked to recommend a good book on a particular subject, to the used-car salesperson selecting a vehicle to match the needs of a customer.

Even within these two examples, the power of a recommender system becomes evident. The librarian, making a judgement based on experience, can enable the visitor to reach the information they need more quickly and directly than if they had sought it by themselves. In a large library with many resources, the librarian becomes a facilitator who can provide shortcuts to knowledge, saving time and energy for the library's users.

However, the used-car salesperson has, of course, less altruistic intentions when directing customers to certain vehicles they suggest are what the customer should be interested in. No longer are the needs of the user the highest priority, as the financial reward in encouraging customers to buy a more expensive vehicle becomes an issue that biases the salesperson paid on commission. The gatekeeper now has an agenda that does not necessarily align with that of those who ask for their recommendations.

The artificial recommendation services used to aid users navigating large information systems also have the knowledge discovery power of the librarian, but can be as biased (whether positively or negatively) as the salesperson. Their utility to users emphasises an alternative, and frequently underestimated information retrieval activity, namely content *discovery* as opposed to the more mainstream content *search*. It is this area of content discovery that can in fact benefit from skewing influences in recommendations, pushing users toward novel content and aid them in browsing media they may not have otherwise come across had the system based recommendations solely on past user behaviour. Artificial recommendation systems, compared to manual systems, also benefit from being more scalable and capable of being applied to datasets found in online media sharing systems.

The field of online recommender systems emerged from work focussing on collaborative filtering systems in the mid 1990s, when systems were developed to improve specific real-world applications like the recommendation of books and CDs in online shops (Linden et al., 2003).

While these initial systems were able to support users in making decisions based on the aggregated behaviour of previous users, they lacked the ability to fine tune the recommendations made to a particular user based on their own specific attributes or behaviour. Any past behaviour that deviated from the emergent trends of the community as a whole was effectively ignored, even when it could have been useful to more finely tailor recommendations.

All recommender systems can be modelled using four fundamental elements:

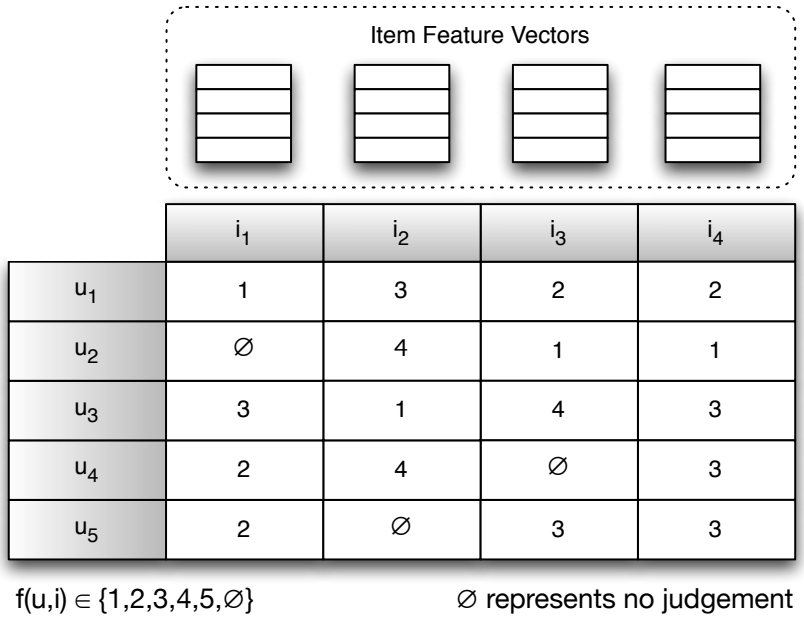
- The set U of users u who interact with the system.
- The set I of items i that form the catalogue from which the system can make recommendations.
- The relevance function f (sometimes called the *utility* function) that maps the rating of a given user for a given item, such that:

$$f : U \times I \rightarrow R \cup \{\oslash\} \tag{2.1}$$

- where R is a totally ordered set of all ratings that the users have given to the items they have rated. Common variations of the nature of the ratings themselves include a finite range of non-negative integers, e.g. Apple's iTunes[7] 5 star ratings, and percentages such as those used by Rotten Tomatoes[8] film scores.

As $R \cup \{\oslash\}$ represents the product of the sets U and I, it lends itself to being visualised as a two dimensional matrix, as shown in Figure 2.3, where each element represents the result of $f(u, i)$ where it exists and $\oslash$ where it does not (signifying when a user has not rated an item). Each $i \in I$ may also be associated with a set of features that describe that item. For example, a film may be summarised by its title, its director, principal actors, etc. Similarly, music tracks could be described by their artist, genre and date of release.

[7] http://www.apple.com/itunes/
[8] http://www.rottentomatoes.com/help_desk/#tomatometer



	i_1	i_2	i_3	i_4
u_1	1	3	2	2
u_2	∅	4	1	1
u_3	3	1	4	3
u_4	2	4	∅	3
u_5	2	∅	3	3

FIGURE 2.3: The set of ratings R for all users U for all items I. Each item has a set of associated descriptive features.

A system can approach the task of making recommendations in two fundamental ways. The first involves analysing the attributes of the items in which the user has previously expressed a positive interest and inferring which other items in the collection most closely match this aggregated profile. The second uses the activity of other people who share similar past behaviour and, using the assumption that those who have agreed on what they find relevant in the past will do so in the future, makes recommendations.

This distinction is recognised in the survey of recommender systems undertaken by Adomavicius and Tuzhilin (2005), where these two approaches are called *content-based* and *collaborative* respectively.

Many existing recommender systems used online exist in one of these two classes, or as a combination of elements of them both as *hybrid* systems. In order to aid disambiguation among other forms of hybridisation defined later in this chapter, I denote this class of hybridisation as *Simple Hybrid*.

2.2.2 Content-based recommendation

By characterising each item in the full catalogue by a set of features, content-based recommenders select items based on the similarity between the set of features from items previously deemed to be relevant and the feature sets of the items in the rest of the catalogue. The capability of these systems is then dependent on two parts: the use of sufficiently descriptive features and an effective method of measuring similarity between feature sets. Considering the potential scale of many online catalogues, which could include millions of items (photo sharing websites for example), a third aspect also becomes important. When many items are judged as being potentially relevant to a given user, these may need to be reduced to a more manageable number. This post-recommendation pruning function also impacts the ability of the system to satisfy the user given a particular interaction scenario.

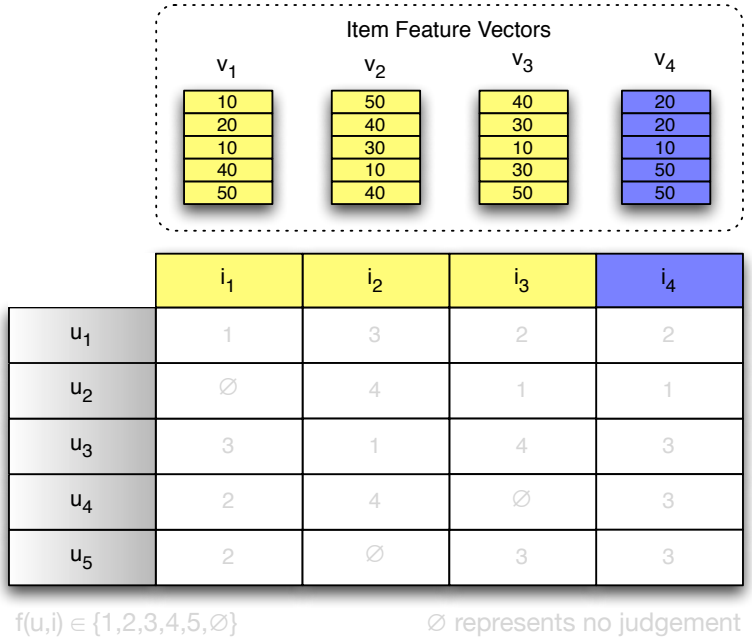


FIGURE 2.4: Using vectors of item feature values as a basis for recommendation.

Figure 2.4 shows how, given an example of an item that has been previously judged as relevant to a user (e.g. it was a book that was previously bought in an online shop, or a music track that was given positive feedback in an online music streaming service), its feature set is then compared to the rest of the catalogue to find the closest matches.

In this example, item i_4 represents a previously chosen item. It is described by a set of 5 features, that for the sake of this example, characterise the item with a value x such that $x \in \mathbb{N}, 0 \leq x \leq 50$. The ordered set of feature values forms a vector. In order to quantify the distance between the vector representing the previously judged item and the feature vectors that characterise the rest of the catalogue, a measure must be chosen. For this example, the cosine similarity of the two vectors both with n elements (denoted $\vec{a}$ and $\vec{b}$) is used, such that a_i is the ith element of vector $\vec{a}$ (and similarly for b_i):

$$\text{similarity}(\vec{a}, \vec{b}) \quad = \quad \frac{\vec{a} \cdot \vec{b}}{\| \vec{a} \| \| \vec{b} \|} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} (a_i)^2} \sqrt{\sum_{i=1}^{n} (b_i)^2}} \tag{2.2}$$

The values found in Table 2.2 are the results of calculating the distance between the source vector and the others.

Table 2.2: Cosine similarity (to 4 d.p.) between source feature vector f_4 and the other feature vectors in the example catalogue.

	v_1	v_2	v_3
v_4	0.9671	0.7606	0.9244

This results in a ordering of items of $i_1 > i_3 > i_2$ in descending feature vector similarity. Using a post-recommendation pruning function of selecting the single highest similar item with respect to feature vector cosine similarity, the final recommendation made to the user would be item i_1.

While cosine similarity is just one example of a distance metric, there are many others. In the work of Hu et al. (2008) the authors evaluated 16 similarity measures and found that, for the image-derived visual features they tested and for tests undertaken on the Corel, Getty and TRECVID2003 datasets, Squared Chord, Fractional ($p = 0.5$), χ^2 and Cityblock distance (Minkowski distance with $p = 1$) measures generally performed better than the other

metrics. Of particular note, they showed how these metrics generally performed better than the simple Euclidean distance (Minkowski distance with $p = 2$), commonly implemented in content-based information retrieval systems.

2.2.3 Collaborative recommendation

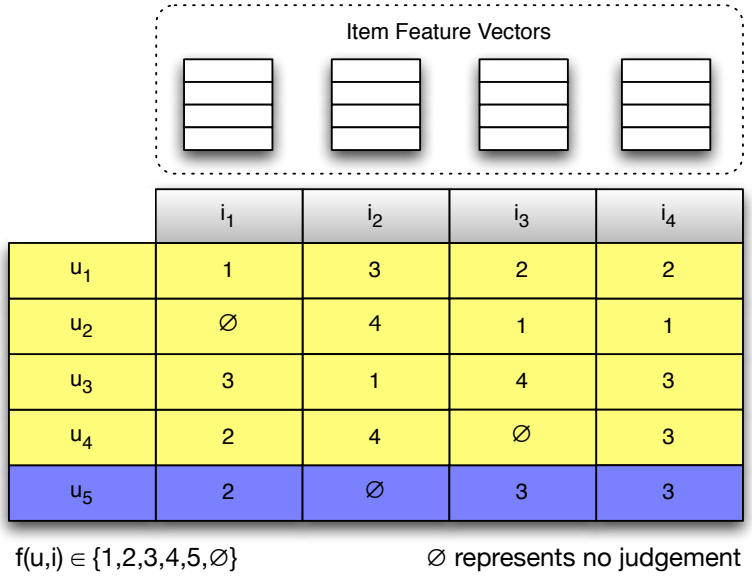


FIGURE 2.5: Using vectors that encode user ratings as a basis for recommendation.

In contrast to content-based approaches that makes recommendations based on similar content, the collaborative-filtering approach makes recommendations based on the behaviour of other users. This technique is used in large online commercial systems, notably online retailer Amazon[9] and film rental service provider Netflix[10] where the judgements and review of catalogue items are used to emphasise new items based on shared past opinions or activity like purchasing.

There are two main techniques in this area: Collaborative Filtering and Market-basket Analysis, both of which are forms of affinity analysis. The former measures the affinity between users and promotes items based on the ratings given by similar users. The latter measures the affinity between the rating information between items and promotes those that have

[9] http://www.amazon.com/
[10] http://www.netflix.com/

frequently co-occurred in some nominal basket, the definition of which varies between applications.

Collaborative Filtering (CF) For a given user requiring item recommendations, CF finds users who are similar with respect to their ratings of items and recommends items from these users that the source user has not yet rated. Using our example catalogue and set of users from the previous section, Figure 2.5 shows how the judgements of user u_5 are compared to the other users in the system.

However, in this case there are elements of the ratings vectors that are empty. In this case, the cosine similarity must be refined:

$$\text{similarity}(\vec{a},\vec{b}) = \frac{\vec{a}\cdot\vec{b}}{\parallel\vec{a}\parallel\parallel\vec{b}\parallel} = \frac{\sum_{i\in I_{ab}}\left(f\left(a,i\right)f\left(b,i\right)\right)}{\sqrt{\sum_{i\in I_{ab}}\left(f(a,i)\right)^2}\sqrt{\sum_{i\in I_{ab}}\left(f(b,i)\right)^2}} \tag{2.3}$$

$$I_{ab} = \{i\in I | f(a,i)\neq\oslash, f(b,i)\neq\oslash\} \tag{2.4}$$

such that $I_{ab}\subset I$ where I_{ab} is the set of all items that both users a and b have rated and that $f(a,i)$ is a rating by user a for item i in ratings vector $\vec{a}$, similarly for $f(b,i)$ and $\vec{b}$. This can be interpreted as calculating the cosine similarity between two vectors such that only items that have a corresponding value in both vectors are included.

Table 2.3 shows this cosine similarity of user u_5 with each user u_x such that $x\in\{1,2,3,4\}$ (see Figure 2.5). Users u_2 and u_4 have the highest cosine similarity and so are deemed to be the user most similar to user u_5. As user u_5 has not judged item i_2 but both u_2 and u_4 have, and those users have been calculated to be the closest to u_5, this item would be recommended.

CF has the advantage of treating each user individually, making their recommendations tailored to their specific profile of previous ratings. However, they must have some previous

judgements in order for similar users to be found and recommendations to be made. Similarly, new items added to the catalogue cannot be recommended to users until they have been rated.

Table 2.3: Cosine similarity between user vector u_5 and the other user vectors in the example catalogue.

	u_1	u_2	u_3	u_4
u_5	0.9949	1.000	0.9872	1.000

Market-basket Analysis (MA) This technique can be considered a kind of complement to CF in that instead of finding similarities between *users'* past behaviour, MA finds similarities among interactions with *items*. In this case, when a user expresses an interest in a particular item—adds it to their basket—the other items in the catalogue that were judged relevant in addition to the chosen item by other users are recommended. This can be done with crude binary ratings such as previously-bought / not-previously-bought in the case of online shopping systems, as well as more comprehensive ratings.

As this technique does not depend on a user's profile of past interactions in order to make recommendations for them, it can be used for users who are anonymous or completely new to the system. However, it also assumes that users who behave similarly share similar interests and makes recommendations based on communal trends and therefore may not match the interests of the user as closely as CF. It also depends on the analogy of a basket, that describes a temporary measure of interest for a given item, which varies between, for example, shopping and photo sharing sites.

Through either CF or MA, recommendation based purely on the judgements of others has the advantage of not requiring analysis of the content of the items in the catalogue from which recommendations are made. However, by analysing the items themselves and incorporating this semantic information into the recommendation process better systems can be made. This is the approach taken by Moshfeghi et al. (2009) in which they use director, genre

and actor information into account in addition to ratings when making film recommendations. In doing so they address the reasoning behind users' interests in particular films, and this is an approach I mirror in my work in Chapter 4 when I use not only the judgements of other users, but also the characteristics of photos themselves in order to predict favourite images.

2.2.4 Hybrid recommendation

The combination of judgements from a community and the correlation of content-based attributes is investigated by Cantador Gutiérrez (2008). In this thesis, the author highlights lack of flexibility of existing collaborative recommender systems to incorporate contextual factors into the recommendation process and emphasises how understanding more about user interests, both implicit and explicit, leads to better systems. In his work the author builds an ontology-based knowledge model which help to overcome the problems inherent in collaborative systems such as user preference sparsity.

2.2.5 Common problems in recommendation

While all three aforementioned classes of recommender (*content-based*, *collaborative* and *simple hybrid*) try to tackle the task of recommendation, they usually suffer from one of two problems, namely data sparsity and bootstrapping.

Data sparsity is a particular problem for collaborative approaches in very large catalogues like those found in online media-sharing systems where only a relatively small proportion of items have been judged and evaluated by other users. Without such judgements, these systems are unable to calculate similarity between new items and those that a user has already judged, and so they cannot be recommended.

The bootstrap problem is the issue of dealing with new users who enter a system and therefore have not built up a set of personal judgements to compare to those of the rest of the

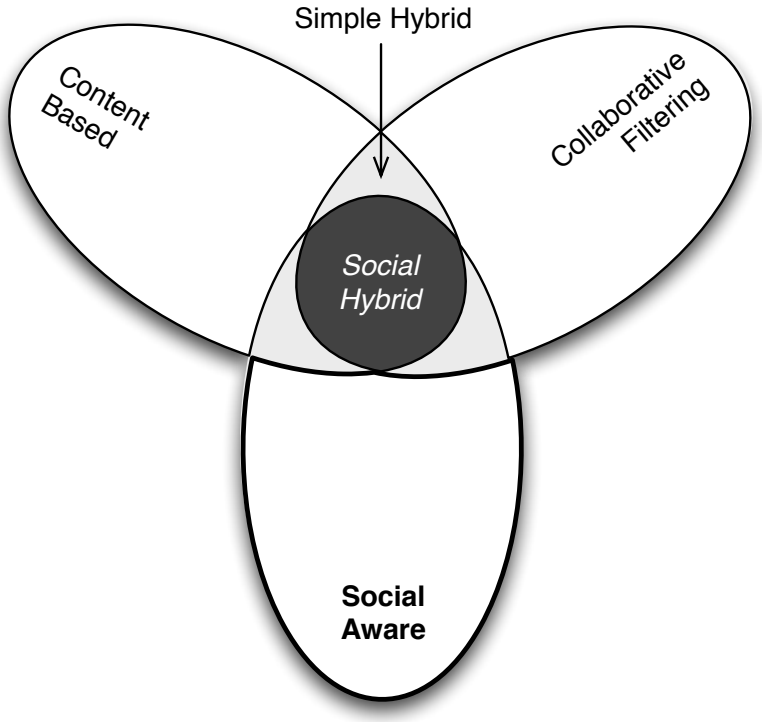


Figure 2.6: Proposed extension to existing recommender paradigms that includes social context information

community. This means that systems reliant on such data (such as content-based approaches that need a number of 'approved' items in order to infer others) may function well for an existing set of users with judgement histories, but will not be able to handle the growth of the community and satisfy the new users that will entail. This can result in different levels of service for different users, with recommendation services only available after sufficient judgement information has been generated. This can limit the quality of service to users who newly join such systems.

2.2.6 New approach proposal

In order to improve on the performance and increase the resilience of existing hybrid recommendation approaches, I propose adding an extra source of data to the process, specifically information that encodes the social connections between users within the community.

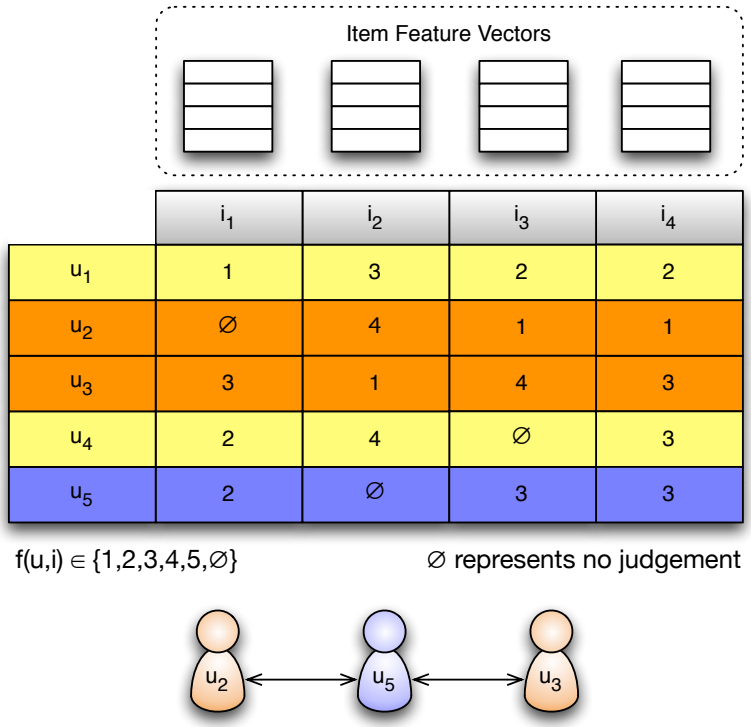


FIGURE 2.7: For a particular user, in this example u_5, recommendations are made not on the judgements of the whole community, but based on that user's social connections within that community

This is analogous to an extension of the hybridisation described previously, as visualised in Figure 2.7, in which the simple social relationship of u_5 with two other users allows the recommendation system to select only the recommendations from that user's network.

Like collaborative filtering, the proposed approach uses judgements from other users to find users who judge in a similar fashion. However, instead of the coarse-grained approach commonly taken that uses judgements from throughout the catalogue, I use personalised, focussed subsets of users. These subsets are defined by the social relationships between the user in question and the other users in the community. As an example, in Chapter 3 this involves subsets of users based on the Flickr contact and shared group membership relations, whereas in Chapter 4 I encode information from these social sub-graphs to characterise image Favourite behaviour among specific local neighbourhoods of users.

This new approach takes advantage of social data available in many media sharing systems, but not currently effectively exploited for the task of recommendation. Even for users new

to the system, if any social context data is available either directly in the system or imported from another, recommendations can still be made. The quality of these recommendations is also likely to improve as judgements start to be made by the user.

It should be noted however that this new approach alone does not help solve the sparsity problem.

2.3 Social context data

2.3.1 Overview of graphs and social graphs in particular

People and the content they generate online are frequently modelled using graph theory, using the paradigm of nodes representing users and/or their media, and edges connecting them. The graphs that result from such modelling lend themselves to comparison with graphs from other kinds of systems, from power grid structure, the neural network of worms and the network induced by the collaboration graph of actors.

It is by such comparisons that discoveries regarding topology have arisen and that go on to influence understanding of other systems. For example, in the seminal paper by Watts and Strogatz (1998) on the dynamics of *small-world* graphs—those in which most nodes are not adjacent but are reachable from most other nodes with very short paths in between—the characteristics of such graphs are modelled and show that they tend to have a node degree distribution that follows a power law.

2.3.2 Social networks online

In work by Cha et al. (2009), Potamias et al. (2009) and Negoescu and Gatica-Perez (2008a), the Flickr social graph based on the contact relationship is shown to be small-world. This implies that most Flickr users are highly-interconnected, are connected to most other users with short chains of mutual contacts and tend to form cliques around hub users. Through

this topological analysis, individual nodes (users) can be characterised and I take advantage of this approach in Chapter 3 where I model tag co-occurrences on photos as graphs and explore them to derive suggestions for other, salient tags.

The Flickr contact graph is also analysed by Kumar et al. (2006) in which they class users based on topological characteristics, namely: *singletons* (lone users unconnected to any others), the *giant component*, made up of users who are ultimately connected to a large fraction of the total graph, for which the average distance between two constituent users decreases as the component grows, and the *middle region* which makes up about one third of users and is comprised of small, isolated communities.

Another work that involved analysing the topology of communities (in this case, scientific author collaboration) was that of Liben-Nowell et al. (2005) who showed how, even without a more comprehensive understanding of users, the general topology of a social graph could be useful in determining future interaction between users. While not able to completely predict the future growth of social graphs (there are many influencing factors that cannot be predetermined), the authors showed that they were able to make good guesses as to future connections between authors (nodes).

In the work of Stoica and Prieur (2009), the authors posited that the different positions of vertices correlate with social 'roles'. They used mobile phone conversations as data, and characterise people with respect to their 'role' of their node in their complete social graph. They label some groups of users as close-knit (a highly-connected subgraph), usually dominated by one node that acts as the centre point of the star, similar to Kumar et al. (2006). They also identify those users who may not be heavily connected to many users, but who do share a strong connection between specific pairs of users. The important ideas to be taken form this work were the recognition of the different roles users undertake in a digitally connected social community, as well as how these roles correspond to real-life social relationships. This particular finding is explored in Chapter 4 in which I investigate the Flickr

Friend, Family (and neither) relations between users and how knowing this piece of real world information makes a difference to understanding the social context of a user online.

An in-depth analysis of a subset of Flickr (users from Spain) was undertaken by Ortega and Aguillo (2008) (in Spanish). They gathered public Flickr data from all those users who self-identified as being based in Spain. The authors showed that this national component of the Flickr social graph grew linearly over the time period they studied (May 2004 to May 2007). They identify a shift in the kind of users within their subset of Flickr during this time frame: early users tended to use the service as a photo storage mechanism, whereas user who joined later were focussed more on the system's social aspects (sharing, interacting with others, etc.). This showed that even in a system not explicitly created to be heavily social, if the mechanisms for users to interact are available, they take advantage of them. This demonstrated a desire to be social and the importance of interaction with other users in a media environment like Flickr.

While many media-sharing systems like Flickr make it easy for researchers to identify social groups by allowing users to affiliate themselves around topics, activities and demographics, this explicit information is not always available. Tantipathananandh et al. (2007) et al. introduce a framework and algorithms to help identify sub-components of social graphs likely to represent social communities. They extend existing approaches to sub-component identification with particular focus on the impact of their evolution over time and the dynamic of the communities they identify. Their work was evaluated on both synthetic and real-world datasets. They highlight the computational complexity of the algorithms they present and stress the importance of a heuristic approach.

The identification of groups within social communities is also investigated by Backstrom et al. (2006) who ask similar questions to those of Tantipathananandh et al. (2007) with respect to community growth over time, but in this case, the authors analyse online communities in LiveJournal[11] and author collaboration networks in DBLP[12]. The former is closer

[11] http://www.livejournal.com/
[12] http://dblp.uni-trier.de/

to the kind of network found in Flickr. They found that affiliation of a user to a group is not simply governed by the number of friends they have already in that group, but also by how those friends are interconnected. This is a finding I exploit in the types of features I use to characterise the social context of a user in Chapter 4.

2.3.3 Analysing groups

Flickr users are able to affiliate themselves into groups, based on themes, activities and attributes of the users themselves. Analysing the value of mining these kind of group membership relations to aid personalisation is not new, in that many systems make recommendations to users based on the content of the groups they have explicitly joined. Teevan et al. (2009) combine the attributes and interaction of users in their approach to "*groupize*" Web search results and identify that their technique works particularly well on explicit groups and group–related search queries. Their work highlights the weaknesses of a one-size-fits-all approach to recommendation/personalisation, even when using more comprehensive data regarding a user and their interactions, including group affiliation information. The recognition of the value of different sources of data and how this varies between users is a phenomenon of which I take advantage when making predictions about Favourite images in Flickr in Chapter 4, where I train models both for communities as a whole and for individual users.

Negoescu (2007); Negoescu and Gatica-Perez (2008b,a) characterise the group participation in Flickr for a sample of the full system, including quantised usage information such as 50.4% of the users in their dataset have at least one photo in a group, confirming that group usage is an important part of Flickr usage for many of its users. They propose five non-exhaustive categories of groups (geographic/event, content, visual style, quality indicator and catch-all) into which groups in Flickr can be placed. In order to summarise the content of a group, the authors use a Probabilistic Latent Semantic Analysis model to determine salient tags that occur in the photos that are submitted to group photo pools. Through

the evaluation of a few hand picked example groups, the authors suggest that the resultant topics are salient. The experiments in this work demonstrate the difficulty in evaluating the relevance of semantic labels to entities such as groups, which usually require manual analysis. This is particularly a problem when judging the capabilities of approaches when subjected to the diversity and scale of data found in web-scale media sharing systems. In order to mitigate this problem when making tag suggestions in Chapter 3, I perturb known data to provide both training and testing data to allow me to perform automated evaluation that mimics manual judgements.

Another issue regarding the approach of Negoescu and Gatica-Perez (2008a) is that their mechanism for gathering data from Flickr only took into account groups to which a user had submitted at least one photo, ignoring those groups of which users were members, but to which they had not submitted photos. This ignores an important section of Flickr group usage that, while different, is still valuable in characterising the interests of users and determining topics for groups.

Negoescu et al. (2009) extend their work to focus on the aggregation of groups by their semantic content. They interpret the problem of group discovery as a clustering problem, and they use a probabilistic affinity propagation algorithm to identify what they call *hypergroups*. Their techniques yields small, homogeneous groups, although again they are limited in their evaluation of these identified groups with respect to semantic correlation and they depend on manual inspection and the proxy metric of the Jensen-Shannon (JS) divergence for group homogeneity (groups that have a small JS value are assumed to be more similar).

One aspect that is common to most, if not all, published literature regarding Flickr social context analysis is the issue of data privacy. The work analysed so far has all used datasets gathered using public data and ignores data that has been designated as private or restricted. This is for two reasons, predominantly because it would be unethical to use data without informed consent but also because it is not possible for researchers (external to the Flickr organisation) to have access to such data in the first place. This means that conclusions

drawn from much work carried out in this field for systems like Flickr that include privacy controls can only be said to hold for a specific subset of the total data available in the system—a point that is rarely, if ever, made by researchers. The excluded private data may well have value in itself, and I argue that by its restricted nature, is considered more valuable by the user. It could therefore provide another source of information that, with the agreement of users, could be used to improve information systems even further.

2.3.4 Social context data

While aforementioned work has focussed on the semantic value of aggregations of media based on social relationships, as well as analysing the topological patterns and structures induced by these relationships, other research has looked at the media itself to recognise social context. One such example is the work of Singla and Weber (2009) in which the authors measure congruence in camera brands between users and find significant correlations between proximity in the Flickr social contacts graph and camera brand. They note the importance of the kind of camera (digital single lens reflex vs. point-and-shoot) being taken into account as well as observing propagation of brand changes throughout the neighbourhoods of subgraphs of "high-cliqueness". The authors' findings are presented as a way of characterising users with respect to their photographic equipment and as a mechanism for measuring change propagation, but I suggest such camera information could also be used to aggregate users to form another implicit graph that connects users based on their attributes/behaviour.

2.4 Text-based feature extraction

2.4.1 Common forms of textual metadata

Many media datasets (a review of which is carried out in Section 2.6) include textual metadata that annotates the content with semantic context. This commonly takes one of two

fundamental forms: either free-text prose or as tags (individual representative keywords or small sets of such keywords).

Prose can range from heavily formulaic text that might be found in data collections that rely on structured retrieval like medical databases (see the work of Olinic et al. (1999) for an example of extending the DICOM 3.0 medical image handling schema with structured text), to unrestrained descriptions provided by users in online sharing environments. Similarly, tags may be used by curators of a dataset within a fixed, defined tag taxonomy with strict rules on their application, or they may be used more freely as part of a *folksonomy* developed by a community of users with shared access to the same media.

The concept of a folksonomy is explored in the work of Mathes (2004) where he attributes the term's coinage (a portmanteau of "tag" and "taxonomy") to Thomas Vander Wal, an information architect who was researching the tagging environments of Flickr and Delicious. In his work, Mathes recognises the diverse nature of the tags used by the community that had formed around these two social media sites. This included identifying tag usage that fell outside of the capabilities of more formal tag taxonomies to encode, such as those tags that had a more functional nature, rather than a descriptive one, like "toRead", a tag that was commonly used in Delicious to mark websites the user wished to bookmark for later reading. He also recognised the emergent nature of the folksonomies that arose from community interaction with their shared media, in that they developed not at the encouragement or enforcement of the systems themselves but naturally through the available interaction mechanisms the users had available.

Mathes also highlighted the benefit folksonomies tend to exhibit when compared to curated taxonomies of being more effective at aiding users when *browsing* as opposed to *searching*, mostly due to the increased coverage the user-generated annotations provided (which is not, therefore, an inherent characteristic of folksonomies, but a byproduct of the mechanism that allows their tags to be widely applied, easily and quickly). Another major benefit of folksonomies when compared to curated taxonomies is how they reflect the vocabulary

of the users who annotate as well as search and browse through their media. This reduces consistency in tag application (leading to ambiguity and an increase in the number of incorrect or low quality tags), but breaks the user away from the single taxonomic definitions of information curators. As an example, folksonomies lend themselves to providing annotations in multiple languages, depending on the languages spoken by the users in the community—usually a challenge for traditional media curation techniques.

Mathes also identifies the class of tags that are used as conversation markers, similar to Huang et al. (2010) as discussed in the following section.

While curated taxonomies have the potential to provide consistent and neutral annotations for media, they can be restrictive, unscalable and rarely reflect the vocabulary of the intended users. Folksonomies, however, are scalable and emerge from media systems that make tagging mechanisms available to users, but are harder to handle effectively.

Schmitz (2006) undertook work to combine the advantages of these two approaches and mitigate their drawbacks. His work focussed on inducing tag ontologies using a subsumption-based approach, which led to a hybrid technique for tag handling. Schmitz highlights two problems with taxonomic annotation models: propensity for low recall and the inability to efficiently or intuitively refine queries. Structured metadata does help users form more structured queries. However, within the context of this particular work, low recall is a weak objection to folksonomy-based tagging models as the author focuses on Flickr and Delicious, systems which have many millions/billions of media items, where high recall is unlikely to be of great importance to users.

With both types, the trade-off between prescriptivism and flexibility leads to a related trade-off between ease of automatic retrieval and the relevance and cost (in terms of manual work) of returned results.

Community-based approaches tackle the issue of cost (in terms of human effort) in labelling media by making it easy for anyone with access to add their own tags. This means it is

possible to tag image collections of the scale that would otherwise require very large teams of professional annotators to manage. It does, however, mean there is less control over the annotation process and can reduce the overall quality of the resultant tags.

The incorporation of user-generated tags into the information retrieval process has been a major focus of recent research. But before its value can be assessed, the reasons and motivations for users in a community to generate such metadata should be explored.

2.4.2 Why use metadata?

Recognising the value in user generated tagging, Ames and Naaman (2007) undertook a qualitative survey of users of both Flickr and a camera-equipped mobile phone photo capture and annotation tool called ZoneTag in order to understand user motivation and what could be done to incentivise more, better tagging. Their taxonomy of tagging motivations (in Flickr/ZoneTag) placed individual reasons to tag along two dimensions: sociality and function. These were then subdivided, as can be seen in Figure 2.8. This recognition that social interaction plays a part in useful tagging behaviour underlies the motivation of the experimental work undertaken in Chapter 3. While Ames and Naaman were principally surveying user motivations, they also made suggestions based on their findings for the effective design of online photo sharing systems. Their last suggestion refers to the value of tag recommendation, as well as warning that they must be aware of the confusing nature of 'inexplicable'[13] tags and that users may be encouraged to add sub-optimally relevant tags if they are presented to them out of convenience.

The first of these problems can be addressed by improving the relevance of the suggested tags and their presentation to users. The second is something that can also be discouraged through a combination of interface design and ensuring updating and improving tags is easy and convenient.

[13] I interpret this in two ways: either the reasoning behind a specific tag suggestion is opaque, or that the tag suggestion is just irrelevant. The authors do not specify.

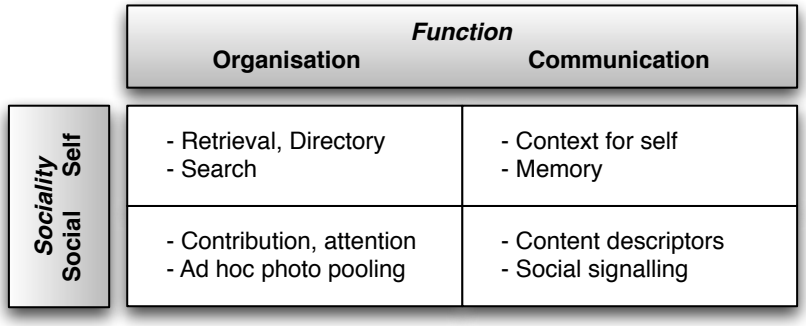


FIGURE 2.8: The dimensions of tagging motivation as proposed by Ames and Naaman (2007)

I tackle this first problem in Chapter 3 where I propose and implement a tag suggestion system using social context data to improve over existing collective approaches.

The two-class classification of tagging behaviour adopted by Ames and Naaman was produced at around the same time as the creation of the Twitter micro-blogging system[14]. The analysis of the *hash tags* used in Twitter (free form tags added to posts to mark topics) carried out by Huang et al. (2010) proposed another factor in classification, namely 'conversational' tags that mark paths through multi-user dialogue. While not regarded as another distinct dimension with respect to those proposed by Ames and Naaman, it does enrich their *sociality* class, reinforcing how important this aspect is when analysing users' tagging behaviour.

In the work of Nov et al. (2008), while recasting their two main classifications as *organisation* and *communication*, the authors study tagging behaviour under the influence of various motivations and social presence situations (both perceived and actual). They conclude that for Flickr, as a specific example of a social online photo-sharing system, making users aware of their social presence (or potential social presence in the community) led to a "positive effect on tagging". This effect seems to consist of an increased number of tags added to a user's photos, which was shown to be generally beneficial for recall and search support by Marlow et al. (2006). It does not, however, directly address the question of tag quality.

[14] http://www.twitter.com/

This is an area addressed by Sen et al. (2007) in their work on tag rating by users. They propose an ensemble learning method (a Bayesian Voting Method) trained on manual tag ratings by users of the MovieLens film recommendation system to select tags to show to users when displaying information about a given film. Their hybrid approach performed well at selecting preferred tags according to their training data. However, I feel that this does not imply that their system was able to select 'good' tags. Their initial definition of 'good' tags are those that "[tie] entities to one another to enhance browsing or search, or may serve as a source of descriptive information". I would suggest that just because a tag is popular or deemed relevant by users doesn't necessarily make it more useful to a retrieval (search) system, even if it enhances user satisfaction while browsing. So while their experimental results do show powerful ways to increase user satisfaction with shown tags, and this valuable, I think their chosen definition of their objective is misleading (or at least ambiguous).

A quantitative metric for tag quality was proposed in the information-theory-based work of Chi and Mytkowicz (2008) in which they measured the information entropy of tags used to annotate postings on the *Delicious*[15] online website bookmark sharing site. They modelled both the tag encoding entropy and the tag retrieval entropy for tags and showed that, as the site became more popular (and the collective tagging vocabulary grew), the effectiveness of tags to discriminate images declined. While they present their approach as a useful technique for objectively measuring the value of a tag to a retrieval system, they acknowledge that their experimental set up was very specific and that they could not provide evidence that their findings would transfer to other, more complex tagging environments. However, they do raise an interesting question when discussing their findings:

> *"What is the balance [between an efficient tag encoding model and an efficient tag retrieval encoding model] and how does it change depending on the information needs of the individuals using any specific social tagging site?"*

[15] http://www.delicious.com/

This reflection on the potentially important influence of personalised approaches to tag recommendation is addressed in Chapter 3 where I purposefully select social subgraphs for tag recommendation for individual users.

2.4.3 Tag handling techniques

The classic vector space model as described in the work of Salton et al. (1975) and first used in the SMART (System for the Mechanical Analysis and Retrieval of Text) Information Retrieval System (Manning et al., 2009) represents documents as vectors of identifiers, quite commonly for text based IR systems as TF·IDF values. This model has been continued to be developed, refined and applied to other, non-textual forms of identifiers like image blobs as shown by Tang and Lewis (2007).

2.4.4 Folksonomies

Folksonomies contrast with manually-curated taxonomies of tags used to annotate images by being generated, not by trained media curators, but by a community of users who share access to tagging mechanisms and the same data. The 'classical' model of a folksonomy consists of a tripartite graph made up of users, tags and resources with relationships forming the edges. This graph can be extended with additional node types as shown in the work of Kern et al. (2008) to other forms of annotation commonly found in photo sharing systems, such as descriptions and comments. In their work, Kern et al. showed how, by extending the folksonomy graph, they were able to provide better tag suggestions in a simulated environment compared to their baseline unenhanced folksonomy based approach.

2.4.5 Tag recommendation

Using folksonomies to improve tag suggestion is a topic also explored in the work of Sigurbjörnsson and van Zwol (2008) in which they use real world Flickr data to model the collective

folksonomy of that system's community and use this to make new tag suggestions in order to support annotation. They assume a task whereby a user is given suggestions based on a small number of existing tags that they wish to extend.

The tagging behaviour of users is analysed and evidence is provided to support the long-held but, until this work, not well substantiated claim that tag usage frequency in Flickr follows a power law distribution. Parameterising this distribution allowed them to prune their potential tag suggestions to ensure they were not either too generic (found in the head of the distribution) or too specific (the tail).

They also highlighted the value of tag recommender systems by showing how little tags are used to annotate photos in Flickr relative to the quantity of media. For example, around 30% of their large (52m photos), representative sample of Flickr of photos that had at least one tag had only one tag. Based on their collective approach they were able to make good recommendations for locations, artefacts and objects.

The authors also showed how effectively an aggregation function could combine different intermediate tag recommendation scores including rank, stability and descriptiveness to form a final value that tags could be ordered by. This (non-personalised) collective aggregation of tag usage throughout Flickr is used as a baseline in the personalised tag recommendation experiments in Chapter 3 where, with Sigurbjörnsson and van Zwol, I extend and refine the approach to use a number of distinct subsets of tags dependant on the social profile of the user.

While Sigurbjörnsson and van Zwol (2008) do not explicitly describe it at such, their tag co-occurrence model is equivalent to a graph-based representation where tags act as nodes and their co-occurrences as edges.

This graph-based approach to tag recommendation (in addition to their FolkRank adaptation of the PageRank algorithm) was compared to standard collaborative filtering techniques (as described in Section 2.2) in the work of Jäschke et al. (2007). By testing their

graph-based approach on three representative datasets they demonstrated the significant performance increase over the collaborative filtering technique according to both precision and recall. This graph-formulation (in this case evaluated on data from Delicious, Last.fm and BibSonomy) is also used in Chapter 3 where I enhance the simple collective graph of tag usage by exploring edges other than simple tag co-occurrence. In contrast to Jäschke et al. my work uses a subset of Flickr as a dataset.

Tag recommendation has also been explored for non-photographic types of media. For example, Sen et al. (2009) develop a film recommendation system ("tagommender") based on a two-stage process that first infers a given user's tag preferences and then uses these tags to infer that user's film preferences. The method of using tags as a metadata proxy for user preferences is a technique I use in my work in Chapter 4 where I use the Flickr Favourite label as a proxy for image preference. In this chapter I also employ both explicit and implicit sources for recommendation, as Sen et al. do in their algorithms for tag inference.

An alternative approach to making recommendation in the same tagging environment is proposed by Zhou et al. (2009) where they use a unified probabilistic matrix factorisation of both tags and explicit user ratings for films. The complexity analysis they undertake highlights a common problem for recommender systems that operate for datasets with millions of items—that of scalability. They show that their approach is a significant improvement in this respect over the existing systems they report. While they do not use implicit information in their approach they, like Sen et al., recognise its potential value to such a recommendation system. I go one step further in Chapter 3 and actively use such data and quantify the improvement to tag recommendation when it is used.

2.5 Feature extraction for image content-based IR

The digital images used in online social media systems are two dimensional representations of photographs that are either based on an underlying discretised matrix of pixel values or a

set of vector drawing instructions. Most are of the former type and, from this point forward, all references to images will be to images of this kind.

In content-based information retrieval, images are commonly represented by extracted features, analogous to how text documents may be described by a set of representative keywords or TF·IDF values. There are a wide range of features that can be extracted and their ability to effectively discriminate between images varies depending on the nature of the image collection they are extracted from and the information need of the user.

The ability of a retrieval system to effectively retrieve images according to a specific information need is then dependent on the effectiveness of the features used to characterise the documents and also the ability of the system to match the information need to the image features.

Such features can be described as existing on a spectrum of 'low' to 'high'—those that characterise images purely in terms of numeric pixel values to those that represent images in terms of more abstract human perception. Low-level features tend to be quicker to compute and simpler to compare, whereas the higher-level features reflect human discriminating ability more accurately and in some cases can be considered to describe the aesthetics of an image.

For example, Datta et al. (2006) look at measuring the aesthetics of images and classifying images as being likely to be rated as 'high' or 'low' by human assessors. They extract nine predominantly low-level features—including saturation and hue, size and aspect ratio and indicators of low depth of field—and use them to train both support vector machines and classification trees when presented with the features of an image. Their work established a quantifiable correlation between visual properties encoded by their chosen 15 features and the assessed aesthetic ratings, yielding an average classification accuracy of 70.12%[16] over

[16] For context, as their datasets were of equal size a trivial probabilistic classifier would be expected to provide an accuracy of 50% across the two classes.

their two classes ("high" and "low") for their particular dataset taken from online photo sharing community *Photo.net*. [17].

Ke et al. (2006) also looked at being able to determine the perceived quality of images. As in the work of Datta et al., the authors derive their chosen image features through *"rules of thumb of photography, common intuition and observed trends in ratings"* from their data collections. These include features that encode colour distribution, hue count, blur and the spatial distribution of edges in an image. Their dataset is also different, coming from the DPChallenge.com[18] website. Their Bayesian approach reaches over 90% precision in ratings images, however this is only achievable for very low levels of recall.

This investigation of the human reaction to and preference for certain images over others is similar to the work I undertake in Chapter 4 where I also use visual features to help judge images with respect to human preference, although I use different features to the two previous citations, as well as a different classification mechanism, and I consider the personal aspect of relevance judgements (not just aggregating ratings from an entire community).

2.5.1 Details of existing features

This section of this Chapter gives a brief overview of main types of features found in the literature, whereas details of the image features used in experiments in this thesis are given in detail in the relevant sections of Chapter 4

There are many visual features that can be extracted from images, commonly split between describing colour, shape and texture information about the scenes they represent. In order to provide a common framework for handling these multiple content descriptors, the MPEG-7 multimedia content description standard was created and later standardised as

[17] http://photo.net/
[18] http://www.dpchallenge.com/

ISO/IEC 15938. This standard provides the tools for media producers and users to effectively annotate media in such a way that enhances both computers', and humans', abilities to exploit the content of media.

While the XML-based MPEG-7 standard provides a framework for describing digital visual information, it is not limited to just this task and is flexible enough to also describe audio, video, 3D models and speech, as well as being used for non-digital media such as printed text.

Regarding the descriptors for visual information, Manjunath et al. (2001) outline the colour and texture MPEG-7 descriptors in their 2001 article. Each of the eight features described (four colour, three texture and one compact texture browsing descriptor) is contextualised and an example method of calculation is given. These core features (see Table 2.4), whilst evidently not an exhaustive selection of all possible descriptors, and not necessarily optimal with respect to effective retrieval, do provide a set of well-defined features and their implementations that allow researchers to compare their systems by using a set of baseline features. They also provide a reasonable trade-off between producing small size descriptors (cheaper to store and compare) that are good quality in terms of discrimination (better at differentiating images).

In extension to the MPEG-7 standardised features, there are others which are also commonly used, which I group here as being based on either texture, colour or shape.

Texture The feature descriptor introduced by Tamura et al. (1978) is an exemplar for those that are based on characterising the textural content of an image. Their descriptor encodes six basic textural features: coarseness, contrast, directionality, line-likeness, regularity and roughness. These correspond with qualities to which the human visual system is particularly sensitive and so makes it a good feature for discriminating between images based on texture in a way similar to humans.

Table 2.4: Overview of a selection of the image content descriptors that make up the MPEG-7 standard.

Name	Type	Brief Description
Dominant Colour	Colour	Salient colours, their percentage in the computed image region and the spatial coherence and variance for a given image
Scalable Colour	Colour	11-bit 256-bin uniformly quantised HSV space histogram that uses Haar transform encoding to reduce descriptor size for greater scalability
Colour Structure Histogram	Colour	M-bin histogram of quantised colours, values representing the count of pixels of given colour that are contained in an 8×8 structuring element as it passes over the image
Texture Browsing	Texture	12 bit description of texture regularity (2 bits), directionality (2×3 bits) and coarseness (2×2 bits)
Homogeneous Texture	Texture	62 value vector comprised of 30 frequency channel energy values, 30 energy deviation values, the mean intensity and standard deviation values for the image
Local Edge Histogram	Shape	240 bit vector of 80 3-bit bins that represent 5 edge direction values for 16 sub-sections of an image

Colour Characterising the wavelengths of light—in other words, the colours—that humans are capable of sensing is a powerful way of describing an image. To this end, extensive research has been carried out on handling the different aspects of this task: quantifying colour, selecting/standardising its range, and representing its value (for a pixel, a sub-image or whole image) in a manner suitable for content-based retrieval. This is frequently undertaken using the concept of a palette of fixed colours that represent a number of visually distinct hues.

For example, the MPEG-7 schema introduced in the previous section includes a description of the Colour and Edge Directivity Descriptor as introduced in the work of Chatzichristofis and Boutalis (2008). In addition to encoding colour information, this descriptor encodes edge and texture information as well.

Colour in the CEDD is initially represented by a 10-bin histogram, with each bin representing a specifically defined preset colour (labelled as black, white, grey, red, orange, yellow,

green, cyan, blue and magenta). The HSV image is split into a preset number of image segment blocks and each is assigned to a bin in the histogram, as decided by a fuzzy rule system that takes each of the three H, S, and V channels as an individual input. The number of blocks in each bin is stored in the feature vector.

An additional four rules applied via a two input (S and V channels) fuzzy system are applied to the 10-bin histogram to produce a 24-bin histogram, which is also stored in the feature vector. These 24 bins represent the number of pixels that have been judged as having one of a Dark, Normal or Light version of the seven colours defined in the 10 bin histogram, in addition to one for each of black, white and grey.

Approaches to tailoring the specific colours a descriptor is most sensitive to introduces a class of colour-based features that focuses on tasks like skin-colour detection. For example, Chai et al. (2003) present their technique that trains a Bayesian classifier to detect pixels in an image likely to represent skin. This is a common approach and there have been assessments of similar approaches undertaken by Vezhnevets et al. (2003), Phung et al. (2005), Singh et al. (2003) and Kakumanu et al. (2007).

Colour features that are tuned to image elements like skin tone are particularly relevant when dealing with images found on online social media sharing sites, where many photos are of people and being able to discriminate between them effectively is important.

Salient points Introduced by Lowe (1999), Scale Invariant Feature Transforms are vectors of local features, generated by a staged filtering process that detects stable points in scale-space. These points are defined as maxima and minima of the result of a difference-of-Gaussian function applied to a series of smoothed and resampled versions of an image. They are invariant to translation, scaling and rotation and partially invariant to changes in lighting and affine or 3D projection. This invariance comes partly from additional blurring

of the image around key locations, mimicking biological processes found in mammalian vision systems. These key points can be represented by histogram descriptors for the points of interest, usually as a 128-dimension vectors.

Lowe proposes using a best-bin-first algorithm (a variant of the kd-tree search algorithm (Bentley, 1975)) to match these descriptors between query and catalogue images as a compromise to the more accurate nearest-neighbour approach as it is quicker to compute—an attribute particularly valued in time-sensitive online retrieval systems.

This basic approach to finding robust features that allow for matching points of interest in images evolved and was extended to make the technique faster and/or cheaper to compute (see PCA-SIFT (Ke and Sukthankar, 2004) and GLOH (Mikolajczyk and Schmid, 2005)).

Inspired by SIFT, Bay et al. (2006) introduced SURF: Speeded Up Robust Features. These features are many times faster to compute than SIFT and the authors suggest that they are more robust to image transforms, more repeatable and more distinctive. Their technique employs an integer approximation of a blob detector based on the determinant of the Hessian matrix (also known as the Monge-Ampère operator) of the image to detect points of interest. These are then characterised by features based on the sum of the Haar wavelet response surrounding the point. Their robustness and computational advantages over SIFT have led to wide-spread adoption in content-based image retrieval systems.

Other image features While the feature classes described so far deal solely with the pixel values that describe the visual content of the image, digital images are frequently annotated with technical metadata encoded by the device that took the image. This can include date and time information, camera parameters, thumbnails version of images as well as textual descriptions and copyright information. Boutell and Luo (2005) explore using this metadata (specifically Exchangeable Image Format - EXIF) to classify images. They were able to use a Bayesian network to effectively class images as to whether they were taken indoors or outdoors, whether they are of a sunset and whether the scene is predominantly

man-made or natural—classes that are useful when handling the kinds of images commonly uploaded to Flickr and Facebook.

2.6 Evaluation datasets for socially-shared multimedia

2.6.1 Introduction

One of the core tenets of the scientific method is the reproducibility of experimental work to allow scientists to evaluate other people's findings. By repeating tests using the same experimental environment and finding commensurate results, conclusions can be confirmed. Individual researchers will carry out repetitions of their own work (repeatability), but for results to be accepted by scientists in the same field, they must be reproducible by others. This requires that it should be possible for all elements of the experimental environment to be replicated externally. In the case of information retrieval (IR) this will include software code (or pseudocode) and data.

Fortunately, most IR experiments do not require exotic or specialist computing hardware, although with the growth in the size of datasets more commonly used in the field, scale has become an issue to the degree that an increasing number of experiments are being carried out on large-scale distributed systems. Software code and algorithms are easily disseminated exactly and completely throughout the community and so this is not a big concern for reproducibility. The biggest issue remaining is the production and dissemination of high quality datasets. Most existing datasets used in the field can be split into either manually-created or user-generated.

Manually-created This includes sets that have been designed specifically for the task of IR evaluation, with carefully selected content, engineered with specific artificial criteria. For content-based experiments, the COREL dataset is perhaps the most well-known and one of the most extensively used and evaluated. Its full set comprises 800 Photo CDs each

containing 100 images grouped by theme. For many early experiments, the full set of 80,000 images was too large and so subsets were selected. Different research groups used different subsets, making comparative evaluation difficult. This problem is explored by Müller et al. (2002) in their 2002 evaluation of the set. Their work demonstrated how easy it was to change the apparent performance of the content-based system without changing the system or even the dataset used. This led to their conclusion that having common access to data is insufficient and that standardised datasets, query sets and corresponding relevance judgements are vital for valid comparison between techniques.

This is a problem that has been addressed within the evaluation fora for visual media retrieval, including CLEF (and specifically ImageCLEF[19]), TREC (Smeaton et al., 2006), specifically TRECVid, ImageEval (currently defunct), MediaEval[20] (formally VideoCLEF) and Benchathlon[21]. Each forum tends to produce and use its own test (and sometime also training) datasets. Their reuse even within the same forum can be limited, making comparison between iterations of the conference difficult.

Community-generated While the aforementioned fora have used manually-created datasets in the past (the BBC Archive rushes set has been used extensively in TRECVid for example), there has been a shift in recent years towards larger datasets derived from online community resources like Flickr, YouTube and the results of search engine image retrieval.

One factor that must be taken into account when producing a dataset is whether the intellectual property rights pertaining to the media (and any derivatives thereof) allow for redistribution and use. With older datasets (COREL for example) there was a single rights holder and clearing the images for distribution was simple. However, with the growth of dataset based on user generated content (for example, MIRFLICKR), there are many thousands of individual rights holders. For this reason, many newer sets only include images which have already been explicitly licensed for use in a research environment, most commonly an

[19] http://www.imageclef.org/
[20] http://www.multimediaeval.org/
[21] http://www.benchathlon.net/

instance of the Creative Commons license. Under the most permissive license (Creative Commons Attribution) the image can be redistributed, and derivatives works made, even commercially, as long as the original creator is credited. More restrictive licenses allow the original creator to have more control over what is done with their media.

Originally, images collected from sources like stock catalogues or news organisation archives had their annotations extracted and bundled along with the images. Some sets started being distributed with a defined set of visual features extracted as well, semantic ontology for tags, etc. This reduced replication of work between different teams working with the sets that were commonly used but independent of the techniques being investigated. More recent datasets have been created from online media sharing sites like Flickr. These involve using a set of queries to query the site and packaging the resultant images, along with their community generated tags, comments and descriptions, as well as any extracted visual features.

2.6.2 Criteria for high-quality social media datasets

While the criteria outlined by Huiskes and Lew (2008a) for datasets aim to address the problem previously mentioned, I propose a variant set of design principles based on their work but that address the specific needs of social media IR, as opposed to multimedia IR in general.

These are grouped into five core areas:

1. Realistic sample of environment Rarely do scientists have the opportunity to measure the entirety of the phenomenon they are investigating (the data found in an entire social media sharing system for example) and so must select a representative sample of data that can be used from which results can be extrapolated. If the sample is characteristic of the phenomenon as a whole, it may be possible for findings drawn from the sample to be induced to hold for the phenomenon. However, ensuring that data that is appropriately sampled from some larger complete set can be difficult for two reasons:

- It may not be possible to evaluate the phenomenon itself to ensure accurate sampling. For example, without knowing the exact full size of an online media collection, gathering a sample of a representative size can be difficult if not impossible. This is frequently the case when access via the programmer's interfaces that many media sharing services provide limits the data they serve to small, focussed subsets (like those in response to text queries).

- For dynamic media collections like community photo-sharing sites, just as with any data sampled from a data source that varies over time, a single static sampled subset of data can only be said to be realistic for the moment it was sampled. This is particularly important for social media datasets as their characteristics change as they grow and evolve, as shown in Section 2.3.

  Conclusions drawn from such a dataset must therefore by explicitly noted as such.

2. Rich media, photos and/or video At the core of a social media dataset is the media itself, either still photos, video or audio, or a combination of all three. These could potentially come from a range of digital systems, and so the media may vary depending on the context of the system from which they come. This context should be recorded and distributed with the media to provide researchers with background information.

3. Contains rich social data The connections between multiple users and between the users and their media are inherently important when working with social media. Therefore it is important for research that wishes to investigate this field fully, that it uses data that comprehensively encodes the complex relationships and interactions between users.

4. Good size For researchers to be confident in their findings, their results need to be deemed significant and so the datasets they use need to big enough for observations to be significant.

5. Availability for repetition In order for the research community to be able to compare findings, the data they use must be shareable and the licenses under which the content and metadata is made available must allow for this. This has been a particular problem for collections based on the Corel and Getty Image databases.

Not all datasets available to researchers are free in terms of cost or in terms of license. "Data as a Service" (Wang et al., 2010) providers are commercial companies that broker access to data conveniently and in forms that data consumers can easily use in cloud computing environments. However, these services frequently charge for their access to data, which can limit accessibility, as well as introducing barriers to sharing specific data samples within the research community. For this reason, while they may make access convenient, they still make less than optimal sources for media for use in research.

Without easy access to stable datasets, experiments cannot be repeated—a fundamental requirement for valid experimentation.

2.6.3 The need for a social media datasets

Non-social media datasets

While media datasets that don't have a social element are evidently inappropriate for *social* media research, they should be acknowledged as datasets commonly used in content-based information retrieval experiments. These sets come from a range of sources, including news agencies (see ImageCLEF 2009 Photo Retrieval Task[22]) and television broadcaster archives (see TRECVID Sound and Vision video files 2007-2009[23]). The IAPR TC-12 (Grubinger et al., 2006) dataset was used in the ImageCLEF photo retrieval task[24] between 2006-2008 and was derived from images provided by a travel company. The previously mentioned COREL dataset would also be considered a non-social media collection.

[22] 498,920 images from Belga News Agency, free-text English captions http://www.imageclef.org/2009/photo/

[23] http://trecvid.nist.gov/trecvid.data.html

[24] http://www.imageclef.org/photodata/

2.6.4 Existing social media and recommendation datasets

In this section I undertake a survey of existing social media and recommendation datasets available to the research community. Each one has been designed with specific experimentation in mind and many have been used in published work. In order to judge whether any of these are suitable for the experimental work carried out in this thesis, I evaluate each with respect to the criteria outlined in Section 2.6.2 and summarise these findings in Table 2.5 found on page 64.

2.6.4.1 Social media datasets

MIRFLICKR-25K

- Developed by Huiskes et al. (2010) and used in their work on visual concept detection and subsequently used by ImageCLEF.
- Comprised of 25,000 Creative Commons-Attribution images from Flickr gathered via public API and include original raw tags, and most EXIF metadata [25] .
- The images were selected to be highly 'interesting' and representative of 'original and high-quality photography'.
- The dataset includes manual annotations using 23 tags (sky, clouds, water, sea, river, lake, people, portrait, male, female, baby, night, plant-life, tree, flower, animals, dog, bird, structures, sunset, indoor, transport, car) as well as single-assessor results (used in the work of Huiskes and Lew (2008a)) as well as results from multiple assessors.
- Intention for it be suitable for relevance feedback benchmarking (Huiskes and Lew, 2008b).
- Also included precomputed visual content descriptors: MPEG-7 Edge Histogram and Homogeneous Texture descriptors and the ISIS Group colour descriptors.

[25] http://press.liacs.nl/mirflickr/

MIRFLICKR-1M

- Extension of MIRFLICKR-25K in that it is a superset of the 25k collection that comprises a total of 1 million Creative Commons-Attribution Flickr images [25] .
- Whilst larger than its predecessor, it only contains manual annotations for the first 25,000 images.

Infochimp - Twitter Census

- *Data as a Service*-based dataset used in the work of Huang et al. (2010).
- As of writing contains 10 million Tweets from 2.7 million users, that exhibit 58 million edges and contain 220,000 *hashtags* (conversation tags) and 2.1 million URLs.

Munmun De Choudhury - Flickr

- This (and the following datasets by De Choudhury) are available from the author[26].
- Contains 2,052 images from 52 Flickr community groups.
- Distributed under Creative Commons Attribution-Noncommercial 3.0 United States License.
- Used in the work of De Choudhury et al. (2009c).

Munmun De Choudhury - YouTube

- Dataset of 11,000 YouTube videos, including tags, views count, comment count ratings, comments authors and comment timestamps.
- Created to study comment threads in media sharing and identifying prominent commenters, see De Choudhury et al. (2009b).
- Distributed under Creative Commons Attribution-Noncommercial 3.0 United States License.

[26] http://www.public.asu.edu/~mdechoud/datasets.html

Munmun De Choudhury - Digg

- Over 151,000 news stories of 56 topics on story discovery site[27], including over 241,000 comments and 94,000 replies, from 10,000 users who share over 56,000 user-user connections.
- 1.1 million *diggs* (personal relevance judgements) submitted between August and November 2008.
- Created to study information diffusion and community evolution, see De Choudhury et al. (2009a).
- Distributed under Creative Commons Attribution-Noncommercial 3.0 United States License.

Munmun De Choudhury - Del.icio.us

- Public website bookmarks from 2,000 users, using 2,000 tags.
- Created to study content popularity and the information roles of users.
- Distributed under the Creative Commons Attribution-Noncommercial 3.0 United States License

2.6.4.2 Recommendation datasets

Book-crossing[28]

- Book recommendation community where users rate books to aid discovery for people with similar literary tastes
- 278,858 users, 1149,780 ratings for 271,379 books

[27] http://www.digg.com/
[28] http://www.informatik.uni-freiburg.de/~cziegler/BX/

Last.fm

- Two datasets provided by Universitat Pompeu Fabra[29] based on data harvested from online music community Last.fm[30]
- 1: 360 thousand users, lists of users' top artists at the time of crawling
- 2: Around 1000 users, full listening histories detailing all music tracks recorded by the last.fm service as having been listened to by each user
- In terms of social data, this set does not capture any relationships between users

The following five datasets are part of the Yahoo! Webscope dataset catalogue, donated by that company for academic use with some restrictions on further distribution.

Yahoo! Music User Ratings of Musical Artists, v1.0

- Musical preference information from Yahoo! Music community
- 10 million ratings of artists gathered over one month period before March 2004.
- Users anonymised but consistently identifiable between ratings
- Designed for research on recommendation systems that use collaborative filter, matrix and graph algorithms (including PCA and its variants) as well as clustering algorithms (Wiyartanti and Kim, 2009).

Yahoo! Music User Ratings of Songs with Artist, Album, and Genre Meta Information

- Also from the Yahoo! Music community, although this collection focuses on users' song preferences
- 717 million ratings of 13 thousand songs given by 1.8 million users.
- Collected between 2002 and 2006.

[29] http://www.dtic.upf.edu/~ocelma/MusicRecommendationDataset/
[30] http://www.last.fm

- Each song has metadata that includes genre information
- Used in the work of Rosenthal et al. (2009) on trust in data labelling systems.

Yahoo! Music ratings for User Selected and Randomly Selected songs, version 1.0

- Ratings for songs taken from two sources:
- 1: Direct interaction from user with the Yahoo! Music website
- 2: Ratings from an online survey undertaken by Yahoo! Research
- In combination, these sources provide ratings from 15,400 users for over one thousand songs.
- Raw data pruned to ensure that each user had at least ten track ratings (derived from web interaction) and exactly ten ratings (from survey) for random songs within the first 5,400 users in dataset
- 300 thousand individual song ratings from web data and 54 thousand ratings from the survey
- Also includes responses from a seven question, multi-choice survey about rating behaviour, given to the first 5400 users.
- Data collected for songs rated in survey were sampled between 22 August 2006 and 7 September 2006.
- Used heavily in the thesis of Marlin (2008) on handling missing data in machine learning, as well as work on collaborative filtering (Marlin et al., 2007) and prediction (Marlin and Zemel, 2009).

Yahoo! Movies User Ratings and Descriptive Content Information, v.1.0

- Film preferences of Yahoo! Movies community, with ratings on scale of A+ to F
- Contains descriptive information about movies released prior to November 2003

- Designed to validate recommendation system based on collaborative filtering algorithms, relational learning, data mining matrix and graph algorithms including PCA as well as clustering algorithms
- Used by Gao et al. (2009) in their work on data fusion and consensus learning.

Yahoo! Delicious Popular URLs and Tags, version 1.0

- 100 thousand URLs stored on the Delicious[31] webpage bookmark management system that had been saved by least 100 users each.
- Contains the ten most commonly used tags applied to each URL, as well as the number of times that each tag was used.
- Designed to investigate tagging behaviour in social bookmarking systems

2.6.5 Dataset evaluation summary

All the datasets so far described are summarised in Table 2.5. For the tag suggestion research found in Chapter 3 all criteria need to be fulfilled except for having ratings. This is needed however for the work on Flickr Favourite prediction in Chapter 4. Licensing is important to both, if I wish to be able to use, publish or redistribute data or subsets thereof. More importantly it means other researchers are able to replicate my experiments with exactly the same data.

It can be seen that no single dataset fulfils all the criteria I have chosen for a dataset to be suitable for the kind of experimentation undertaken in the following two experimental chapters.

[31] http://www.delicious.com/

Table 2.5: Overview of existing social media and recommendation datasets available to researchers as presented in Section 2.6.4. Legend: ? = unknown, ✓= available, ✗= unavailable, — = not applicable, m = million, k = thousand

Dataset	Number of users	Number of items	User interaction with media?	Interaction between users?	License	Ratings?	Media type
Social Media Datasets							
MIRFLICKR-25K	?	25k	tagging	—	CC-A	✗	photos
MIRFLICKR-1M	?	1m	tagging	—	CC-A	—	photos
Infochimp—Twitter Census	2.7m	10m	(hash)tagging	✓	?	—	tweets
MdC—Flickr	?	2,052	tagging	✓	CC-AN	—	photos
MdC—YouTube	?	11k	tags and comments	(via comments)	CC-AN	(comments)	videos
MdC—Digg	18k	151k	comments	✓	CC-AN	✓	news stories
MdC—Delicious	2k	?	tags	✗	CC-AN	—	websites (URLs)
Recommendation Datasets							
Book-Crossing	278,858	271,379	ratings	✗	?	✓	books
Last.fm	360k/1k	?	play count	✗	?	✓	music tracks
LY! Music—Artists	?	?	ratings	✗	?	✓	music tracks
Y! Music—Artist, Album and Genre	1.8m	13k	ratings	✗	?	✓	music tracks
Y! Music—Selected Songs	15,4000	71k	ratings	✗	?	✓	music tracks
Y! Movies	?	?	ratings	✗	?	✓	films
! Delicious URLs and Tags	?	100k	tags	✗	?	✓	websites (URLs)

2.7 Reflection on the state of the art

This chapter has introduced the field of digital media and recommender systems that operate in this environment. I have shown that existing systems focus on one of two main ways of making recommendations and have proposed a new extension to this paradigm that takes advantage of increasingly available data about users and the interactions within their communities—their *social context*.

I have presented and evaluated the current state of the art with respect to handling textual, visual and social information and shown how the latter has not been fully exploited in research concerned with recommendation associated with online digital media.

Data used by previous researchers is also evaluated according to criteria I propose characterise suitability for experimentation undertaken in the following two chapters. I show that while there are a number of social media and recommender system datasets available to the community, none match my requirements and justifies my decision to design, create and evaluate my own.

Chapter 3

Personalised Tag Suggestion Using Social Context

> *"The intelligence of that creature known as a crowd is the square root of the number of people in it."*
>
> (Terry Pratchett, Jingo)

Roadmap In this chapter the field of metadata recommendation and its value in systems that handle large scale online media collections is introduced. I highlight current shortcomings of existing implementations and propose a framework that takes advantage of social context data to improve such systems. Through experimentation I undertook in collaboration with Drs. van Zwol and Sigurbjörnsson, I show how current methods can be improved significantly and my findings are evaluated with respect to online demographic breakdown. An initial study is presented that demonstrated the feasibility of the approach, as well as more comprehensive experiment that addressed realistic use-case scenarios for a wide range of potential users.

Note: This chapter is based on work that was predominantly carried out while working in the laboratory of Yahoo! Research Barcelona with Dr Roelof van Zwol and Dr Börkur Sigurbjörnsson. As such, some sections of this chapter closely reflect the content of the papers that were published on this work, particularly our 2010 *RIAO* paper (Rae et al., 2010).

3.1 Motivation

To address the first and second of the research sub-questions in Section 1.2.2, an experiment is described in this chapter that investigates the value of data derived from different types of social interactions by using them in the pre-existing use-case of tag suggestion for media annotation. These systems are found in large-scale image sharing environments, where they suggest tags to users, either when they are annotating their media, or when they are searching and browsing and need help refining their queries. Adopting a previously investigated use-case like tag suggestion allows for comparisons to existing work that make it easier to demonstrate any potential value of the novel kinds of features being used to augment the system.

Tagging of media objects has been shown (as discussed in the Chapter 2) to be a powerful mechanism that can improve search options for images and video in social media sharing sites such as Flickr[1] and YouTube[2]. Agreement in tagging tends to emerge where people agree on the semantic description of a media object.

In popular social media sharing sites there are billions of images and videos being annotated by millions of users and a crowd-sourced vocabulary gradually forms based on their descriptions. These tags provide a wealth of information that can form the basis of recommender systems (see Section 2.2 for overview).

[1] http://www.flickr.com/
[2] http://www.youtube.com/

Ten years into the new millennium, few[3] systems make use of such data other than tag occurrence and co-occurrence values. While Sigurbjörnsson and van Zwol (2008) have shown that this kind of data can be used effectively, the performance of such systems are still limited in three ways:

Quality of source data

Crowd-sourced data derived from non-experts can lead to inconsistent data. The consequent recommendations based on this data can hence be of lower quality.

Scope of source data

Using only tag-based statistics as an indicator of tag recommendation quality ignores the other influences concerning suggestion relevance.

Catering for lowest common denominator

Mining data from throughout the whole community and looking for general trends in tag usage means that results are not tailored to specific users. Users vary in their attributes and online behaviour. By being more selective in which data are used to train a recommender system, individual user behaviour can be taken into account and final performance increased.

While the quality of tags derived from crowd-sourcing is unlikely to approach that of those from expert annotators, there is scope for improvement by being selective in choosing tags to use for recommender system training. For example, by pruning out tags that are evidently inappropriate (machine tags, tags used by very few people, etc.) and emphasising those with high consensus or social relevance to particular users, recommendations based on this source data will improve.

[3] At the time of writing, many sites including Facebook, Picasa Web and Photobucket do not offer users suggestions for tags when annotating media. While it is difficult or impossible to know how the internal workings of some commercial web-based enterprises, those that do offer tag suggestion do so in a way that suggests only rudimentary underlying mechanisms.

3.2 Tag suggestion

3.2.1 Problem specification

Recommender systems can be useful in an online media sharing environment like Flickr in a number of ways. For example, a user annotating a photo that they have uploaded to Flickr can be recommended tags related to the photo that can be used to extend any existing annotation. This automated suggestion helps to simplify the task of expanding the coverage of the tags that describe the image and hence increase the ability of the system to accurately retrieve the image given a user's information need. In another scenario, the recommender system can provide recommendations while searching. This can be done through automated query expansion, or in an interactive process by means of search assistants that provide additional query terms that the user can choose to add to their query. To build such systems, training data is required from which to derive potential suggested tags.

Typically systems based on 'collective knowledge' aggregate the annotations used in a large collection of media objects independently of the users that defined the annotations, an example of which is found in the paper of Sigurbjörnsson and van Zwol (2008). Alternatively, the recommendations can be personalised by using the annotations for the photos of a single user as in the paper of Garg and Weber (2008). Both approaches come with their advantages and drawbacks. When the recommendations are based on collective knowledge the system can make good recommendations on a broad range of topics, but is likely to miss some recommendations that are particularly relevant to specific users. Basing the recommendations on the personal data of a user will provide good results if the user has been actively using the media sharing system, making the statistics underlying the recommendation system reliable, and if the user is conscientious while annotating. However, if that is not the case, the system will have trouble trying to make recommendations.

Therefore an approach that merges information from both collective and more personal aspects of data could outperform systems that use only one.

3.2.2 Social features

Users participating in social media sharing sites interact with others, as visualised in Figure 3.1. For example, in Flickr, users can maintain *contact* relations with other users, who then can be further identified to be their *friend*, *family* member, or *other* type of contact. There are also less explicit connections that can form between users, based either on indirect interactions like being members of the same interest group on Flickr.

	Direct	**Indirect**
Explicit	Contact "Friend" "Family"	Group membership
Implicit	Age Location	Comments on same photo

FIGURE 3.1: Flickr social features can be generally classified along two axes: directness and explicitness.

The group membership of a user defines the explicit interest of a user in a certain topic, or community of users sharing a common interest. By taking these affiliative social connections into account, a fuller picture of the users' interests can be modelled.

3.2.3 Tag suggestion using social context

To address the three main problems with existing systems highlighted at the beginning of this chapter, a system is outlined here that aggregates and exploits the information from four different contextual layers of the social network that exists among Flickr users, in an extendable probabilistic framework. As the approach focuses on the individual user, the first user-specific contextual layer is the "Personal Context" (PC), constructed from the annotations provided by the user. Second, a "Social Contact Context" (SCC) is defined by aggregating the annotations from all users that are identified as a contact of that user. Third, a "Social Group Context" (SGC) is obtained by aggregating the photo annotations of photos

posted in the groups that the user is subscribed to. Finally, a "Collective Context" (CC) is derived by aggregating the annotations for all photos posted by all users.

A tag co-occurrence graph is derived for each context, based on analysis of tags used to annotate the photos within that context. Different vocabularies and co-occurrence statistics emerge per user for each of the four contexts.

The Personal Context is derived from a user's personal tag dictionary. I propose that it is likely to be more accurate than the Collective Context when recommending tags but likely to have smaller coverage. The social activities of a user are of great influence on the size of the Social Contact and Group Contexts and so their relative performance is dependent on the scale of these activities. As the effectiveness of these two contextual layers has not been studied before in similar recommender systems, the evaluation of this experiment focuses on these two contextual layers. The recommendations based on the unpersonalised Collective Context are used as a baseline for comparing the quality of the suggestions from the individual personalised contexts.

3.3 A probabilistic approach using social graphs

3.3.1 Probabilistic prediction framework

This section uses terms and methods from mathematical graph and set theory. For an introduction to graph theory sufficient to cover the requirements of this section, I recommend 'Graph Theory' by Reinhard Diestel (Diestel, 2006).

Definitions:

Occurrence A tag 'occurs' if it is used to annotate a photo.

Co-occurrence Two tags 'co-occur' if they have been used to annotate the same photo.

For each context a co-occurrence tag multigraph is derived for its set of constituent photos, with nodes representing unique tags $t_i \in T$ that annotate photos in the set and edges occurring when two tags have been used to annotate the same photo. This means that when two tags co-occur multiple times, there are multiple edges between the their two tag nodes in the graph. As the concept of 'co-occurrence' is inherently undirected, so too are the edges that represent these relations. The potential for multiple edges between nodes also implies that the final graphs are not *simple*, and as tags cannot co-occur with themselves, the graphs cannot have *loops*.

The occurrence tally $o_F(t_i)$ and the co-occurrence frequency $c_F(t_i, t_j)$ can be calculated for all the tags of all the photos in the set F.

The conditional probability of one tag occurring in a photo given the fact that another tag co-occurs with it in the same photo within set of photos F is formulated as:

$$p_F(t_i|t_j) = \frac{c_F(t_i, t_j)}{o_F(t_j)} \tag{3.1}$$

To produce a set of recommendations for a given set of input query tags, each query tag is first used to generate a intermediate set of recommendations and these sets are then combined. The intermediate set of recommendations S for a given query tag in a given context is the complete set of tags s that co-occur with that tag, i.e. are adjacent in that context's graph. The final recommendations are emphasised in terms of their rank position by penalising those tags that are not recommended by all query tags.

So, to calculate a set of recommendations given a set Q of input query tags, the probability of an intermediate suggestion given Q in context x where $x \in \{PC, SCC, SGC, CC\}$ is first calculated for each tag s:

$$p_x(s|Q) := p_x(s) \prod_{q \in Q} max\left\{p_x(s|q), \varepsilon\right\} \tag{3.2}$$

where ε is a non-zero value significantly smaller than the lowest conditional probability in the complete set of all conditional probabilities. This value is introduced because in cases where recommended tags do not co-occur with all input query tags, any instance of non co-occurrence would reduce to zero the overall probability of the recommended tag given the query tags.

This would mean that in a list of output suggestions ordered by descending conditional probability, there would be a pool of tags at the bottom with the same value of zero. It is more desirable to keep these tags low down in the final output list of suggestions, but maintain some sense of ordering based on the tags that do co-occur (in case the top N elements of the list that we wish to return to the user include some of these tags that do not co-occur with all the query tags), and the use of ε achieves this.

For this particular experiment a value was derived by using the maximum tag occurrence value as given in Sigurbjörnsson and van Zwol (2008) of around 12,500, which would give a minimum conditional tag probability of $\frac{1}{12{,}500} = 0.00008$ assuming this highly popular tag co-occurred with a tag that only appeared once. By reducing this value by a factor of around 10, the final value of $\varepsilon = 0.00001$ was used.

Each resultant probability $p_x(s \in S|Q)$ is used to produce an ordered list of tags in descending order of probability. The top N tags are then the final recommendations as given by that context's network of tags for a given query tag set, where N is the number of tags that best suits the use-case scenario of the recommender system. This method can be used in an identical manner for any similarly structured graph of tag co-occurrences.

3.3.2 Personal Context (PC)

The personal set of tags for a given user is made up of all the tags used on all the images that the user has uploaded. These sets vary between users, but consist solely of information relevant to that particular user. These sets tend to be far smaller and less comprehensive than that of the general tag cloud discussed in Section 3.3.5, but better reflect a user's personal ontology of keywords, or *personomy* (Jäschke et al., 2007).

It is this user-specific nature of the Personal Context that I suggest allows it to make highly relevant recommendations to specific users.

3.3.3 Social Contact Context (SCC)

A user in Flickr can explicitly connect themselves to other users by giving them the label 'Contact'. These inter-personal connections form a social graph between the users in the system where users are represented as nodes, and a directed edge can exist between two users if one labels the other as a 'Contact' (an edge in the other direction can exist when the connection is reciprocated), with an upper limit of 3,000 [4] non-reciprocal relationships per user.

I produce a tag co-occurrence graph from this data by taking all the photos from the contacts of the user for whom recommendations are being generated and aggregating them. This excludes the photos of the user themselves. This means that there is no overlap in the photos between the Personal Context and the Social Contact Context, making it easier to evaluate the contribution of the tag graph based on the photos derived from the 'Contact' label in isolation.

The tags suggested using this network capture the vocabulary not of the user but of their online social community, possibly sharing attributes like language, geographical proximity

[4] This upper limit is dictated by the Flickr infrastructure.

and to some degree photographic interests, which would be helpful in providing a more precise set of recommendations.

3.3.4 Social Group Context (SGC)

Users on Flickr can interact with each other by becoming members of shared interest groups and sharing photos with others who have done the same. There are therefore images associated with such groups and the tags annotating these images may have a common theme—the topic of the group. These group topics vary immensely, from visual themes (e.g., black and white, High Dynamic Range) to subject themes (e.g., landscape, portraiture) and activities (e.g., A Photo A Day, reportage of real word events). These topics can be very wide and vague ("Nature") or very specific (photos from a particular real world event). The Social Group Context aggregates the tags of the photos associated with the groups of which a user is a member to form another tag network that can also be used to derive possible tags for recommendation. These recommendations should more closely represent the interests of the user in terms of the photos they interact with as opposed to their attributes, better described by the Social Contact Context.

3.3.5 Collective Context (CC)

Whereas the previously defined tag graphs have been selected subsets of the entire collection of photos available in Flickr to better reflect certain aspects of the user requiring recommendations, the Collective Context aggregates the tags from all photos from all users. This forms a very large tag graph that encapsulates the tag usage of the whole community.

While it is not user specific, it does provide an extensive dataset from which to make recommendations. It also has the advantage of being able to provide recommendations when the user is not very socially active (i.e. has few contacts or is not a member of many groups, etc.) which would restrict the capacity of the personalised contexts to provide relevant results.

3.3.6 Tag co-occurrence multigraph definitions

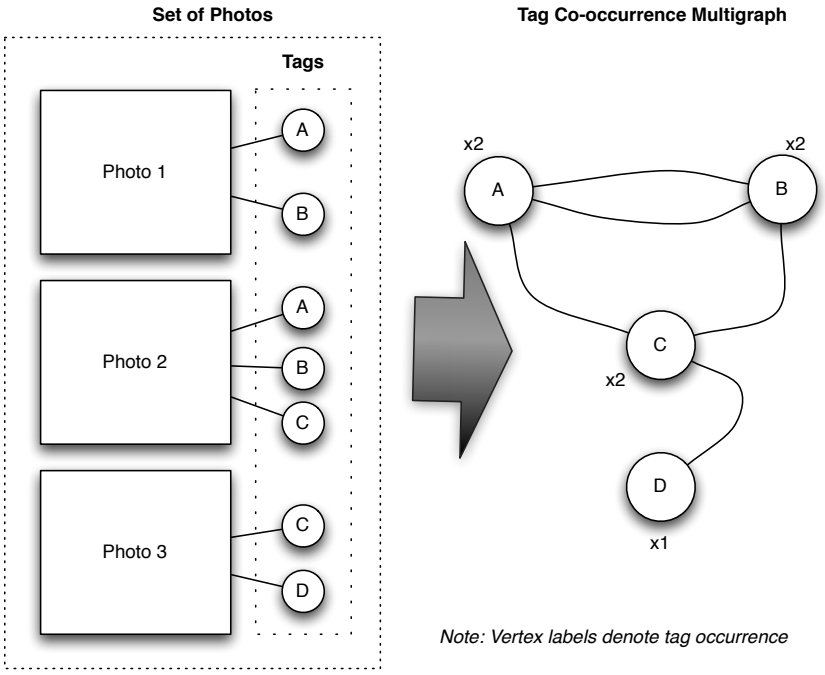


FIGURE 3.2: Example of the vertex labelled multigraph induced by a set of photos and their tags

Tag co-occurrence multigraphs are formulated for a set of photos, as visualised in Figure 3.2

The multigraphs for all four contexts defined above are defined as follows:

Notation:

- Let P be the set of all photos p in Flickr.
- Let T be the set of all tags t used to annotate photos in Flickr.
- Let w_p be the set of all tags t that annotate a photo p.
- Let c_p be the set of co-occurrences of the tags of photo p defined as:

$$c_p = \{\{x, y\} \,|x, y \in w_p, x \neq y\} \tag{3.3}$$

- Let C be the multiset of all co-occurrences of tags of photos in Flickr such that:

$$C = \biguplus_{p \in P} c_p \tag{3.4}$$

where the multiplicity of a tag pair is equal to their co-occurrence within Flickr.

Note: I use the symbol $\uplus$ to denote the multiset sum, as used[5] in Syropoulos' "Mathematics of Multisets" (Syropoulos, 2001).

Definition 1: The induced multiset of tag co-occurrences of a set of photos Similarly, given any set F of photos, the function $\varphi(F)$ is defined as:

$$\varphi : F \mapsto \biguplus_{p \in F} c_p \tag{3.5}$$

Definition 2: The induced aggregated set of tags of a set of photos Given a set of photos F, the function $\psi(F)$ is defined as:

$$\psi : F \mapsto \bigcup_{p \in F} w_p \tag{3.6}$$

Definition 3: The Collective Multigraph The vertex labelled tag co-occurrence multigraph W derived from the Collective Context is induced by the tags in Flickr as the nodes and the co-occurrences among those tags as the edge relations thus:

$$W = (T, C) \tag{3.7}$$

with a vertex labelling $l : T \mapsto L$ where $l(t) = o(t)$.

[5] There is a minor typographical error in the author's description of multiset commutativity in Definition 7, part i), which should say that $\mathcal{A} \uplus \mathcal{B} = \mathcal{B} \uplus \mathcal{A}$ and not $\mathcal{A} \uplus \mathcal{B} = \mathcal{B} \uplus \mathcal{B}$ as stated.

Comment.: The following multigraphs can all be denoted by the general form $G = (T, \varphi(F))$, where G is the sub-multigraph of the Collective Graph W induced by the tags attached to the photos in F.

Notation:

- Let U be the set of all Flickr users u.
- Let y_u be the set of photos uploaded by a user u.

Definition 4: The Personal Multigraph The tag co-occurrence multigraph X_u induced by the Personal Context for a user u is then defined as:

$$X_u = \big(\psi(y_u), \varphi(y_u)\big) \tag{3.8}$$

Notation:

- Let D_u be the subset of U that are the contacts of user u.
- The set of photos of the contacts of user u is then defined as:

$$Q_u = \bigcup_{v \in D_u} y_v \tag{3.9}$$

Definition 5: The Social Contacts Multigraph The tag co-occurrence multigraph Y_u induced by the Social Contact Context of user u is then denoted:

$$Y_u = \big(\psi(Q_u), \varphi(Q_u)\big) \tag{3.10}$$

Notation:

- Let G be the multiset of all Flickr groups[6], such that a Flickr group g is a set of photos.
- Let h_u be the submultiset of G that is the multiset of Flickr groups that user u is a member of.
- The multiset of co-occurrences of the tags of the photos in the Flickr groups of user u is then:

$$C'_u = \biguplus_{g \in h_u} \varphi(g) \tag{3.11}$$

- The set of tag nodes induced by the same photos is defined as:

$$T'_u = \bigcup_{g \in h_u} \psi(g) \tag{3.12}$$

Definition 6: The Social Group Multigraph The tag co-occurrence multigraph Z_u induced by the Social Group Context of user u is denoted:

$$Z_u = (T'_u, C'_u) \tag{3.13}$$

Generalisation of Graph Model

This graph-based formulation of aspects of Flickr can be used to model other similarly interconnected systems. All the graphs described in this chapter have at their core photos that are annotated with tags. The personalised graphs are interconnected with edges based on social relationships. This could be made more generic: resources that have descriptors, interconnected through general interaction.

By doing so, the model described here could be fitted to more diverse systems. For example, it would be very easy to model documents in a digital library, annotated with keywords, and

[6] The term *group* used in this Section refers to the Flickr concept of user groups, and should not be interpreted as any kind of mathematical set.

that are used by multiple users who group themselves by institution, work role, interest, etc. Resources would be recommended that suit the task or role of a particular user, using the same recommendation framework as presented in this Chapter.

Similarly, the graph-based model presented here could be used as an alternative to existing methods of product recommendation used by supermarkets (see Section2.2). The model would describe shopping trips (analogous to *photos*) that involve buying products (*tags*) from a supermarket (*Flickr*), by multiple shoppers (*users*) described by demographic attributes. Recommendations could be made for products based on the various ties between shoppers; shared demographic attribute, behaviour, etc.

All systems that could be modelled using the graph-based formulation from this chapter would be able to take advantage of the same probabilistic prediction framework presented in this thesis.

3.3.7 Aggregation methods

In order to maximise performance, the four individual ordered lists produced from the tag networks described previously are combined. A number of methods for rank combination (also called *data fusion*) that are prevalent in recent literature, including rank concatenation using an ordered hierarchy (which I call Fall Back), linear combination based on score or rank values (see Frank Hsu and Taksa (2005) for overview and evaluation) and machine learning algorithms (Lee, 1997; Bartell et al., 1994; Burges et al., 2005).

This section describes and compares these options.

Fall Back The four input ranks are ordered from most user-specific to least: personal, social contacts, social group and then collective, as shown in Figure 3.3. For a given number n of required suggestions, a new rank is constructed from a copy of the most personal rank available of length p (where p is the length of the complete list of all possible suggestions made by that context). Where it does not provide at least n

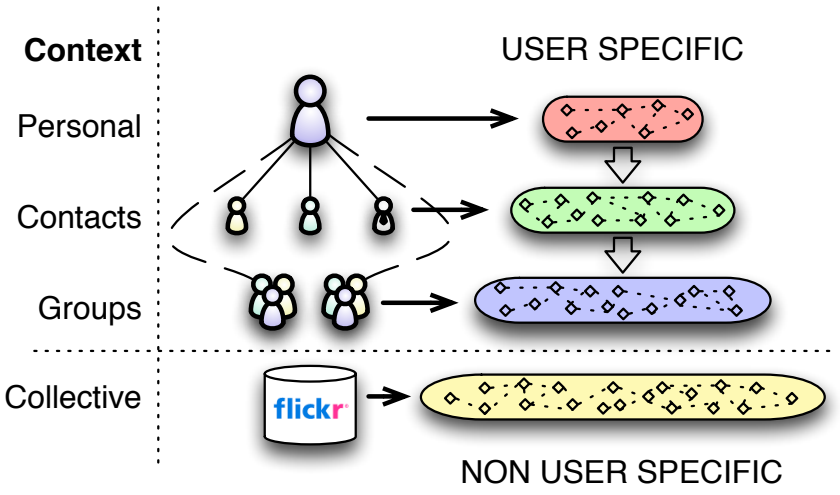


FIGURE 3.3: Hierarchical ordering of contexts going from most personalised to most general.

suggestions, the first $n - p$ highest scoring suggestions from the next most personal rank are appended onto the end of the ranking, ignoring duplicates. This is repeated for each context until n suggestions have been gathered.

Borda Count This is a group consensus function that combines voting ranks by assigning descending consecutive integer scores to each element of the individual ranks and summing values (or averaging when all elements are common to all ranks) to produce a new ordered rank, as described in the work of Van Erp and Schomaker (2000). This means that when a suggestion occurs in only a subset of the total set of ranks, it will get a lower score than if it had occurred in all of them at a similar rank. This emphasises suggestions that are made by more than one rank.

The basic Borda Count method treats each input rank equally by not weighting them and uses linear scoring. There are issues when dealing with ranks of differing lengths as this method is based on the assumption of additive independence, which is not fully justified in this case. For example, the top score suggestion from one rank may be considerably worse than the top score suggestion from another, but they would be treated as equivalently good suggestions by this implementation of the Borda Count method.

This method emphasises those suggestions that are common to more than one constituent rank and can therefore also penalise relevant suggestions that were on produced by a single input rank.

In my implementation, the scores assigned to the ordered ranks start with the first element of each rank being given the same value equal to the length of the longest of all the input ranks.

Modified Borda Count This is based on the basic Borda Count (above) but the starting score for each rank is equal to the length of that rank. This emphasises those ranks with greater recall.

Linear Combination of confidence values Summation of min-max normalised tag probability values given the query tags, with the maximum and minimum values calculated from all the tag probabilities used by all the ranks being combined.

Multi-Layer Perceptron Taking a different approach to the previous methods, the problem of rank combination can be cast as a classification problem, whereby a model is trained to judge whether the suggestions made by the four constituent ranks were relevant or not by assigning a binary relevance label to each of the training examples. The relevance label was determined to be *true* when the suggestion occurred among the prediction set for that photo, and *false* otherwise. The final set of suggestions from the test set were those that the classifier judged as relevant.

The ranks were fed into the Weka (Witten and Frank, 2005) implementation of a Multi-Layer Perceptron, a classifier that uses back propagation to classify instances. The MLP network had sigmoidal nodes and 44 hidden layers (42 example attributes + 2 classes). The MLP has a learning rate of 0.3, a momentum of 0.2 and was trained in 500 epochs and these values are kept the same for all instances of training. I recognise that these parameter values are not optimal–the task of deriving the parameters that yield the highest performance is left to future work.

For each example of a user with a query tag set in the data collection, I produced a set of suggestion tags from each of our four contexts. I trained the MLP classifier using 85% of the total examples and tested on the remaining 15%. These training examples were described by 42 features split up into 4 groups. These were:

1. the independent probability of the two query tags and candidate tags occurring in the personal and general contexts, giving 12 values;
2. the conditional probability of each query tag with each candidate tag for each context, giving 16 values;
3. the probabilities of a candidate or query tag divided by the conditional probability of a candidate or query tag given another candidate or query tag, giving 8 values;
4. user specific statistics including the number of contacts the user has, the number of groups they affiliate themselves with, the number of photos they have submitted to Flickr and the dictionary size of their personal context of tags, giving 6 values.

I found that a cost-sensitive meta classifier boosted performance by weighting the importance of returning true positive results from the MLP classifier. The optimal value of penalising misclassification of relevant examples 9 times more than irrelevant ones was determined by exhaustively testing increasing values between 1 and 15 (where numbers higher than 15 were deemed high enough to no longer change performance), and choosing the one that gave the highest performing output with respect to early precision and mean average precision.

After testing the aforementioned rank aggregation techniques, I found that the Borda Count provided the best performance with respect to my chosen metrics. For this reason, I chose this method for the experiment.

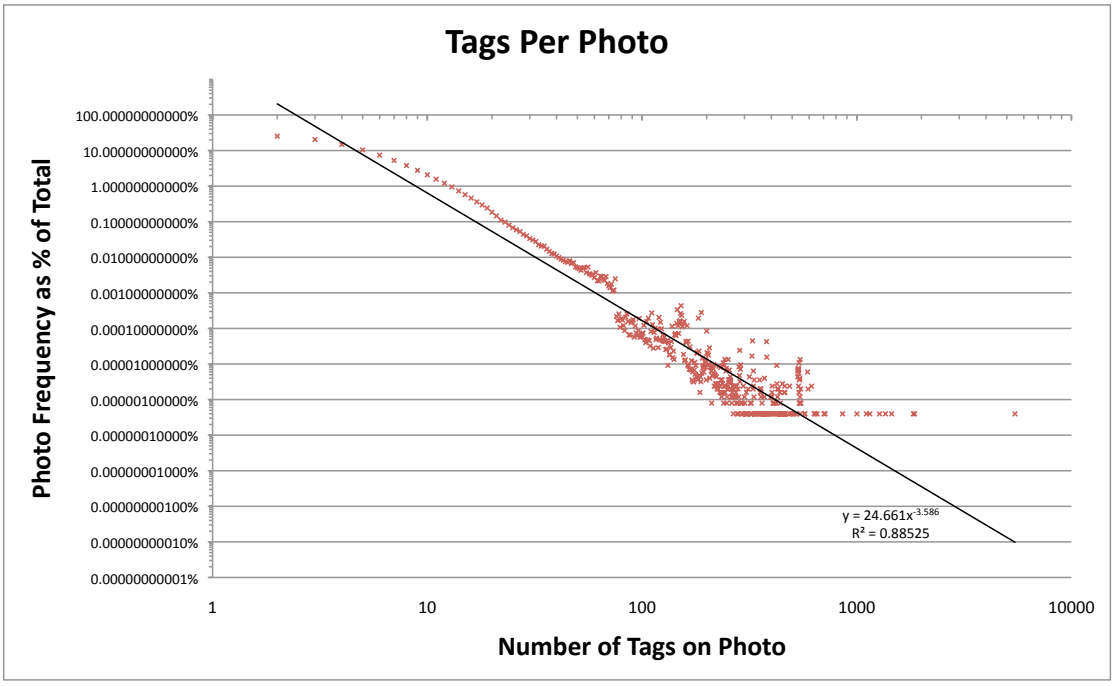


FIGURE 3.4: The distribution of tags per photo for the subset of Flickr photos that have two or more tags, as of May 2008, which results in a total of 250 million photos.

3.3.8 Data processing

Figure 3.4 shows the distribution of tags per photo, where the number of tags is greater than two, for all Flickr images as of May 2008, as a percentage of the total dataset. This distribution is is characterised by a power law trend line of $f(x) = 24.661x^{-3.586}$ with a coefficient of determination R^2 of 0.88525, as generated by Microsoft Excel 2008 for Mac.

With respect to calculating tag co-occurrences however, the salient distribution trend is a little different. For a short time after Flickr was first launched until 30th January 2007[7], photos could be annotated with an unlimited number of tags. After this date however, a limit of 75 tags per photo was imposed and the majority of photos in Flickr have been subject to this limit and, as of writing, this is still the case.

Figure 3.5 shows the distribution of tags per photo, where the number of tags is greater than two (at least two tags are required for a tag co-occurrence) and less than 75 (the current maximum number of tags a user can annotate a photo with) for all Flickr images as of May

[7] http://blog.flickr.net/en/2007/01/30/news-2007-1-30/

2008. The removal of photos with more than 75 tags induced a reduction of $1.06 \times 10^{-4}\%$ (to 6 d.p.) in the number of photos in the dataset. The resultant power law trend line to fit this distribution is then given by $f(x) = 21.882x^{-3.264}$ with and R^2 error of 0.9775.

This trend line is a closer fit to the data according to its R^2 value, which implies the distribution of tags per photo for photos with more than 75 tags is less consistent with that of those with fewer than 75 tags.

In order to give an idea of the scale of co-occurrences calculations required in the experiments described in this chapter, I assume the following:

- The number of tag co-occurrences for a photo that has s tags is given by:

$$g(s) = \left(\frac{s^2 - s}{2}\right) \tag{3.14}$$

- The number of tags per photo is distributed such that the proportion of photos in a sample of Flickr that have x tags is given by $f(x) = 21.882x^{-3.264}$. This is based on the power law trend line calculated for the graph in Figure 3.5 for photos that have between 2 and 75 tags.

- The total number of co-occurrences, as a proportion of the size of a sample of photos, is then the bounded area under the distribution trend line, the integral:

$$\int_2^{75} f(g(x)) = \int_2^{75} 21.882 \left(\frac{x^2 - x}{2}\right)^{-3.264} \tag{3.15}$$

  However, as $f(x)$ is a discrete function (the number of tags attached to an image cannot be fractional), this simplifies to:

$$\sum_{x=2}^{75} f(g(x)) = \sum_{x=2}^{75} 21.882 \left(\frac{x^2 - x}{2}\right)^{-3.264} = 22.5685 \tag{3.16}$$

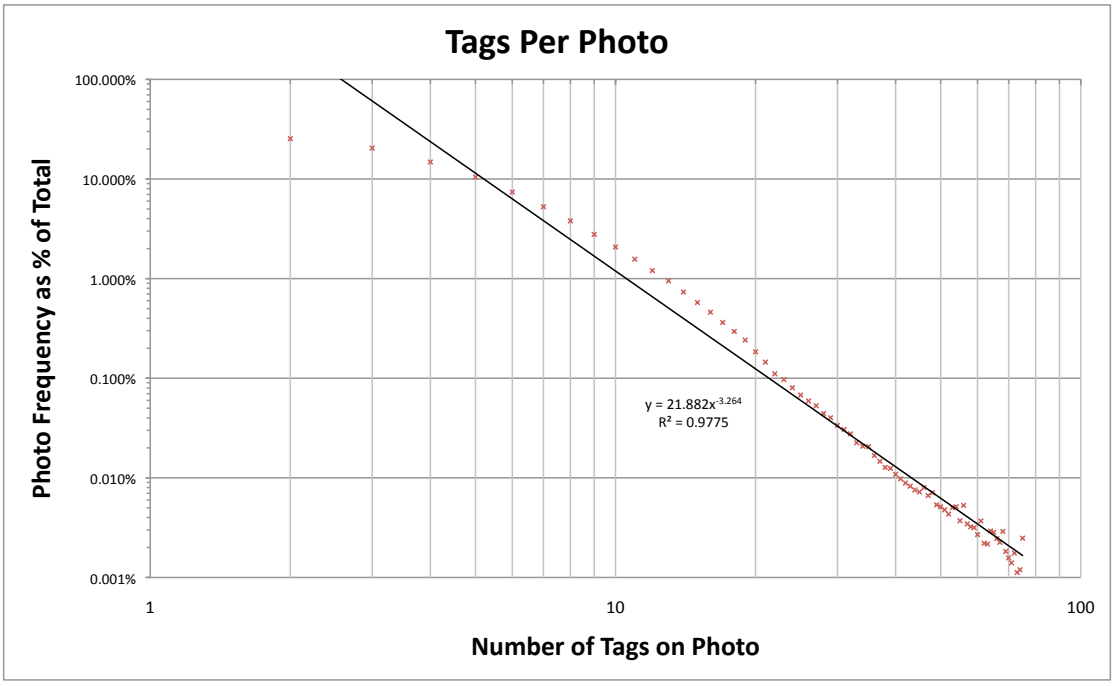


FIGURE 3.5: The distribution of tags per photo for the subset of Flickr photos that have between 2 and 75 tags inclusively, as of May 2008, which results in $1.06 \times 10^{-4}\%$ fewer photos than the complete set of 250 million photos.

So, for a sample of 250 million photos that conform to the preceding assumptions, the total number of co-occurrences is given by:

$$250,000,000 \times \sum_{x=2}^{75} 21.882 \left(\frac{x^2 - x}{2} \right)^{-3.264} = 5.683 \times 10^9 \tag{3.17}$$

This implies that the number of co-occurrences increases linearly with respect to photos if distributions assumptions hold, with a scaling factor of 22.5685.

The graphs derived from the Flickr tag co-occurrence are particularly large (e.g. 250 million nodes induces over 5.6 billion edges) due to the heavy interconnectedness caused by the co-occurrence of tags on photos (see Equation 3.14). The computation and processing of these tag graphs for web-scale datasets like those used in this chapter quickly becomes infeasible with respect to time and space using standard data processing techniques on a single computer.

The Data Management Group[8] at Universitat Politècnica de Catalunya, Spain, developed a prototype graph-based database system DEX[9]. DEX (Martínez-Bazan et al., 2007) is specially designed to perform well on tasks common to the processing of graphs, such as retrieving data according to topological constraints (e.g. returning data according to graph nodal adjacency), making it well suited to handling the kind of data for this experiment. However, DEX was only (at the time this chapter's experiments were undertaken) capable of running on a single machine and so with its inherent limitations on fast memory, DEX was not capable of handling the induced graphs.

Hadoop[10] is a system architecture that stores data on a large, shared filesystem that allows an array of low-cost nodes to share processing, using the Map/Reduce computation paradigm (Dean and Ghemawat, 2008).

The processing itself is directed by a series of Map/Reduce programmes implemented in Java, that break the required processing down into multiple simple steps that can be repeated over the entire large data collection. Hadoop is ultimately capable of calculating the required tag probabilities for this experiment within a feasible time frame (hours instead of days or weeks) and is more robustly implemented than DEX. However, the method of incremental calculation of tag probabilities means that intermediate stages create lots of very large files, particularly for the SGC networks where the size of the tag probability lists ranged from tens of megabytes into the hundreds of gigabytes, depending on the user.

In order to make calculation feasible, a compromise had to be made in the probabilistic framework. Whereas ideally all tags that are connected to the query tags within the network would be evaluated as potential candidates for suggestion, a subset was selected to reduce computation. This subset was chosen using a Candidate Selection Process, outlined

[8] http://www.dama.upc.edu/
[9] http://www.dama.upc.edu/technology-transfer/dex/
[10] http://hadoop.apache.org/

in Algorithm 1, parameterised with K, which is the number of candidates to return for the set of query tag.[11]

Through exhaustive parameter exploration, a value of $K = 10$ was deemed to balance the need for graphs of manageable sizes that were still capable of making relevant suggestions.

Input: Co-occurrence graph $G = (T, E)$
Input: Q, a non-empty set of input (query) tags
Input: K, candidate compromise cut-off parameter
Output: $C\prime$, a set of output (candidate) tags
begin
  $C = \emptyset$;
  $L = \emptyset$;
  foreach $q_i \in Q$ **do**
    $C = C \cup \{t | t \sim q_i, t \in T\}$;
  end
  $L =$ set of sets l_{t_i} such that $t_i \in C$, l_{t_i} is a set of tuples of form <tag, probability>
  foreach $\{\{t_A, t_B\} | t_A, t_B \in C, t_A \neq t_B\}$ **do**
    if $t_A \in Q$ **then**
      $p = P(t_A|t_B)$;
      $l_{t_A} = l_{t_A} \cup < t_B, p >$;
    else if $t_B \in Q$ **then**
      $p = P(t_B|t_A)$;
      $l_{t_B} = l_{t_B} \cup < t_A, p >$;
  end
  sort all $l_{t_i} \in L$ by descending probability;
  $C\prime$ = first K elements for each $l_{t_i} \in L$;
end

Algorithm 1: Pseudo-code description of the candidate selection process

During the feasibility testing stage of the experiment, techniques for managing Hadoop tasks continued to develop both internally to Yahoo! where I was undertaking the experiment and within the wider community. Hadoop users realised that much of the processing commonly undertaken on map/reduce clusters was quite similar, and so higher-level scripting languages were developed to reduce the burden on users. This meant that by the time the feasibility study had concluded, new methods were available to implement the tag probability calculations that, due to improved parallelisation and sub-task management, were more computationally efficient with respect to the data processing undertaken in this experiment.

[11] In the ideal situation where computational resources and time were not restrained, the full conditional probabilities as defined at the beginning of this Chapter would be equivalent to using $K = \infty$.

The MySQL-like query language PIG[12] (Olston et al., 2008) saw a particularly fast rate of development and take up by research scientists using Hadoop, and it was determined that the experiment could be redeveloped using this language.

This had two main advantages: code that consisted of hundreds of lines of complex data manipulation could be reduce to a few tens of lines of this higher-level language; and the compromise regarding the selection of candidate tags could be eliminated due to sophisticated parallelisation and task handling techniques used by the PIG library. This is reflected in the boost in performance with respect to my metrics seen during the evaluation of the main experiment.

3.3.9 Experiment strategy

Having seen from related work how extensive the tag networks for users and their communities can be, there was doubt as to whether processing such a large amount of data would be feasible using a standard laboratory desktop computer. To measure the feasibility of implementing such a tag recommendation system and determining how scalable the data processing would be, I designed an initial study with Drs. van Zwol and Sigurbjörnsson. This study used the same simulated task as the main experiment, but used a far smaller number of users—in this case 25—and a more space-efficient variation of the probabilistic framework, as described in Section 3.3.8. After completing this feasibility study, the conclusions drawn from its results and evaluation of its implementation, informed the design of the main experiment. Additionally, advances in cluster based processing techniques meant that the original, more-accurate probabilistic framework could be implemented without compromise.

[12] http://hadoop.apache.org/pig/

3.4 Experiment design

3.4.1 Task

Figure 3.6 shows the distribution of tags from all Flickr photos as of May 2008 that had 2 or more tags, a set of 250 million photos. 61% of the photos that match this criteria had exactly two tags, the remaining 39% with three or more. If the tag suggestion system presented in this chapter could make it easier to annotate this large percentage of photos in Flickr, the added metadata would make it both easier and quicker for users to find relevant images.

The performance of the system is therefore evaluated through a "proxy task" - for a given photo with 10 tags or more, two tags are taken as input for the system and its performance is measured in terms of how many of the photo's remaining tags it can recommend. This partitioning of a photos tags is shown in Figure 3.7.

Since this is a "simulated" evaluation of the system, it may not give the correct picture of the absolute performance of the system. In fact, it is likely to underestimate the performance of our system as the metrics used only take into account exact matches. The system may be capable of producing tags that are relevant for a particular photo but are not in its prediction set. These results are currently ignored in this evaluation. However, the consistent approach taken is appropriate for comparing the relative performance of different tag recommendation methods.

To address this issue of under-reporting performance, I propose using human evaluation of tag suggestions to judge relevance using an online crowd sourced method. However, this was not possible during the work presented here due to time and cost, and the lack of available tools at the time of experimentation. This is something that could be addressed in future work.

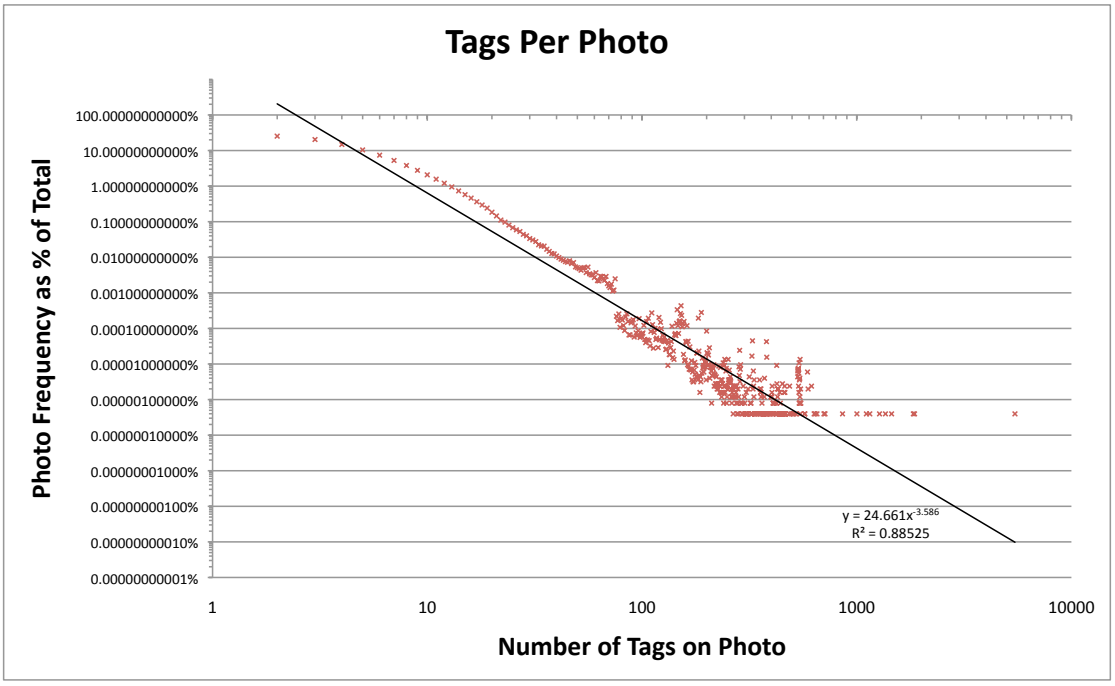


FIGURE 3.6: The distribution of tags per photo for the subset of Flickr photos that have two or more tags, as of May 2008, which results in a total of 250 million photos.

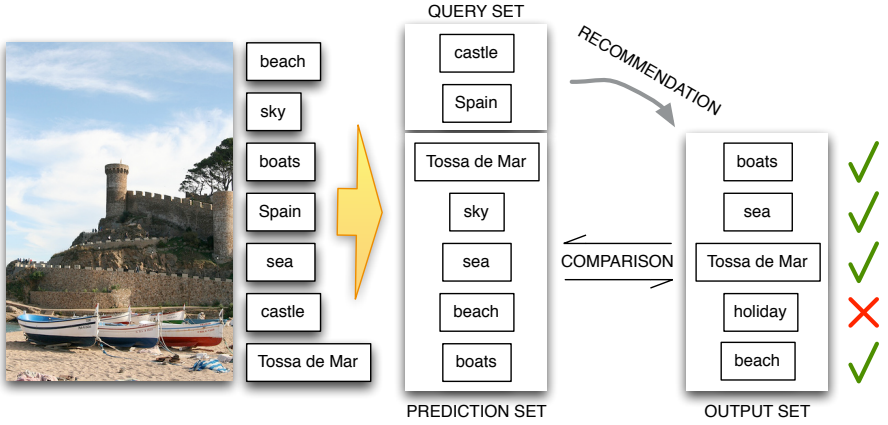


FIGURE 3.7: Example of partitioning of existing photo annotations to provide source and test data.

3.4.2 Input tag selection

As shown in Figure 3.6, photos that have exactly two tags represent 61% of photos that have two or more tags in Flickr—a significant, well defined subset of photos of Flickr to be able to address and potentially augment with more tags with a recommender system. In this task two tags are selected from those that currently annotate a photo (reflecting the case of a photo with only two tags) to produce the query set. The rest become the prediction set

(which reflects a set of potential tag suggestions). The choice of which two tags are used as input is important, as not all tags are as likely to generate tags that match the prediction set.

FIGURE 3.8: The all time most popular tags on Flickr as of 11th October 2010.

Figure 3.8 shows a tag cloud of the most popular tags used in Flickr since its inception, those that are shown larger having been used more often. The most popular of these tags include "wedding", "party" and "nature" which are, I propose, tags of general specificity. Tags which are more general are likely to have more connections to other tags in a tag network than more specific ones.

In traditional Information Retrieval terms, general tags could be thought of as having good recall, but lacking in precision. Conversely, more narrowly specific tags are more likely to provide a smaller number of more relevant suggestions—higher precision, but lower recall.

Therefore there is a trade-off between those that provide a wide range of suggestions and those that provide a few highly relevant ones. The concepts of term specificity and exhaustivity are explored in the work of Sparck Jones (1972) where the author demonstrates the

value in taking into account tag frequency within a collection when weighting terms, to balance these two dependant attributes.

I propose three potential tag selection methods:

Random Two query tags are selected at random from among the image tags. This method makes no assumptions about the ordering of tags as given by the user, and treats them all equally.

Order of addition The first k of total n tags added by the user are selected. This method assumes that users approach tagging with a strategy or ordering in mind and that tags can be attributed to a position on a continuous scale of general to specific. I propose, based on personal observation of tagging behaviour in Flickr, that the first tags added to a photo are more specific than the last $n - k$ tags and could therefore be better candidates for providing highly relevant tag suggestions.

Most specific An external tag evaluation function could be applied that is trained to select the most specific tags, regardless of the order of addition. Such approaches have been taken in the work of Sparck Jones (1972).

3.4.3 Evaluation considerations

The priorities in the objective analysis of the performance of my system are directly related to the envisaged interaction scenario on which the experiment is based. This scenario is analogous to the traditional information retrieval task that considers the set of query tags to be the query and the prediction set to be the resultant rank of "relevant documents".

I suggest that users do not want to be swamped with many low-relevance tags. It therefore important to ensure that the few tags that are suggested are as relevant as possible.

To calculate the performance of the system in producing the prediction set I use the *trec_eval* tool[13]. I measure the performance using standard information retrieval metrics for ranked

[13] http://trec.nist.gov/trec_eval/

retrieval experiments: Precision of the top n recommended tags (P@n), Mean Reciprocal Rank (MRR), and Mean Average Precision (MAP).

P@n is precision as calculated above with respect to only the top n ranked retrieved elements. In this study I use P@5, as users are unlikely to want to process many more than around 5 tag suggestions when deciding which are relevant for their media.

$$Precision = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|} \tag{3.18}$$

The reciprocal rank of an ordered list of tags is the multiplicative inverse of the rank of the first relevant tag. The average of this value for multiple suggestion tasks is the Mean Reciprocal Rank. Users do not want to have to evaluate many suggestions and this metric quantifies how high up relevant results occur in the list of tag suggestions.

$$MRR = \frac{1}{|Q|} \sum_{i \in Q} \frac{1}{\text{rank}_i} \tag{3.19}$$

where Q is the set of relevant tags that are found in the output rank of tag suggestions and rank_i is the ordinal rank position number of tag i.

Mean Average Precision measures the average of precision values computed at each relevant entry in the rank of suggested tags. Like MRR it also emphasises relevant results at high rank positions. Using the definition given by Manning et al. (2009), the set of relevant tags for a query set $q_j \in Q$ is $\{d_1, ...d_m\}$ and R_{jk} is the set of ranked retrieval results from the top result until arriving at document d_k, then

$$\text{MAP}(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} \text{Precision}(R_{jk}) \tag{3.20}$$

When a relevant document is not retrieved at all, the precision value in the above equation is taken to be 0.

It should be noted that while the tag suggestion system presented in this chapter might recommend tags relevant to the photo, the above metrics will only take into account the exact matches with the target tags, as specified by the user. The final values generated by these experiments must therefore be interpreted within the particular context of the experiment and should only be seen as indicative measures of performance when compared to other research outside of this thesis.

All results were tested for statistical significance at p-value levels of 0.05 and 0.01 using the Student's T-test, as this has been found to be reliable for this type of information retrieval experiment (Sanderson and Zobel, 2005). All significance tests are performed relative to the baseline of that part of the evaluation.

3.5 First stage: feasibility study

The experimental task for this initial study tests how feasible (in terms of both space and time) it is to generate the different types of tag contexts (personalised, social and collective) and to combine them to form a system that can be used to help users annotate their photos. These contexts can potentially contain many thousands of tags for each user, and so discovering whether it is possible to extract and use these tag graphs in a reasonable time frame and using a feasible amount of storage is important. This is especially true if any resultant system that emerges from this experiment is to be used in an online environment where spontaneous data processing and presentation is key. This feasibility study is conducted on a small test sample of users and its evaluation is used to inform the experimental design decisions of the full experiment.

3.5.1 Dataset and users

The tag dictionary used in this feasibility study contains the annotations of over 250 million public Flickr photos that had two or more tags, uploaded before May 2008 and available

Statistic	Min.	Max.	Mean	StDev	Median
No. Contacts	4	88	49.5	27.5	61
No. Groups	1	429	83.4	91.7	63
No. Photos	259	4686	919.0	1056.6	609

Table 3.1: Statistics over the 25 users in our experiment.

through the Flickr API[14]. The tag dictionary is limited to tags from this collection applied to images by at least 5 users. This decision was based on analysis by Schmitz (2006) that showed that this cut-off level reduced the number of unhelpful tags in the tag set considerably, without overly compromising the comprehensiveness of the set. The resulting tag vocabulary contains roughly a million unique tags.

25 Flickr profiles were selected to form my set of test users. This number provided a manageable collection of users for whom tag suggestions could be calculated without onerous computation time within the constraints of the processing infrastructure available. They were selected at random from among users that represented a variety of "socialness"—i.e., the collection contained users with few contacts and users with many contacts to better allow for the observation of how this factor affects the ability of the system to make suggestions. Table 3.1 shows some characteristics of the 25 users in terms of the number of contacts they have, the number of groups to which they belong and the number of photos they have in the dataset.

For each of the users, two sets of photos were collected for evaluation—those uploaded before and after May 2008. The earlier set was used to produce the tag networks that generate tag suggestions, and the latter provided test examples. A total of 250 photos, 10 for each user, were collected for use with all aggregation methods. Only photos with at least 10 tags were selected. For each photo in the test data, two tags were randomly selected to produce a *query set*, while the rest of the tags for that photo became the *prediction set*.

Due to the scale of the dataset used, a distributed, parallelised approach was taken to process the tag occurrence and co-occurrence values required. This was done on a Hadoop[15] cluster

[14] http://www.flickr.com/services/api/
[15] http://hadoop.apache.org/

using 100 nodes, with processing taking a few hours, using the manually implemented Java-based Map/Reduce implementation outlined in Section 3.3.8. This processing produced all the conditional probabilities for all users in the dataset.

The system took each query set of tags as input and returned a list of recommended tags according to our probabilistic framework.

3.5.2 Evaluation of results

While this feasibility study is primarily designed to test the practicality of processing the required quantity of data for the framework I propose in this chapter, I undertake a short evaluation of the results to produce early indicators of system performance that I can use to compare to the output from the full experiment.

First the performance of the framework is evaluated using different contexts in isolation, and subsequently in combination. Different baselines are used for these two stages of the evaluation. For the first stage, the Collective Context is used since it is comparable to the system presented in Sigurbjörnsson and van Zwol (2008) and is non-personalised. This makes it easier to examine the effect of personalisation.

In addition to analysing the relative performance of the individual contexts, this evaluation also includes analysing the combination of the individual runs to measure how this affects performance. This involves looking at the results that come from using the multiple combination methods described in Section 3.3.7. The baseline for the combination runs is similar to that presented in Garg and Weber (2008). For clarity, I report only those that give the best performance according to my chosen metrics.

Run	**MRR**	% difference from CC	**P@5**	% difference from CC	**MAP**	% difference from CC
Collective Context (CC)	0.2665		0.1040		0.0642	
Personal Context	0.2065	-23%	0.1144	10%	0.0709	10%
Social Contacts Context	0.0959	-64%‡	0.0352	-66%‡	0.0151	-76%‡
Social Groups Context	0.2692	1%	0.1128	8%	0.0650	1%

TABLE 3.2: Feasibility Study Experimental Results
Individual context runs. MAP values marked with ‡ are statistically significant with $p < 0.01$

3.5.2.1 Performance of individual Contexts

The results of evaluating the different contexts in isolation are shown in Table 3.2. Neither the Personal nor the Social Groups context gave statistically significant differences in performance when compared to the Collective Context and so conclusions cannot be drawn regarding their relative performance.

The Social Contact Context performs consistently badly, compared to the collective baseline for all three metrics.

I observed that the Social Group Context tended to perform in a similar manner to the Collective Context. This could be because it makes use of the collective knowledge of a large set of more diversely annotated photos. However, I suggest it could be more focused than the Collective Context since it addresses a set of photos that I suggest are closer to the user's photographic interest, based on the themes or topics of the groups they are involved in.

It is worth noting that the runs looking at just one context at a time, by themselves, do not have statistically significant improvement over the baseline. In some cases however, they do contribute to the significant results of the combinations of contexts, as shown in the next section.

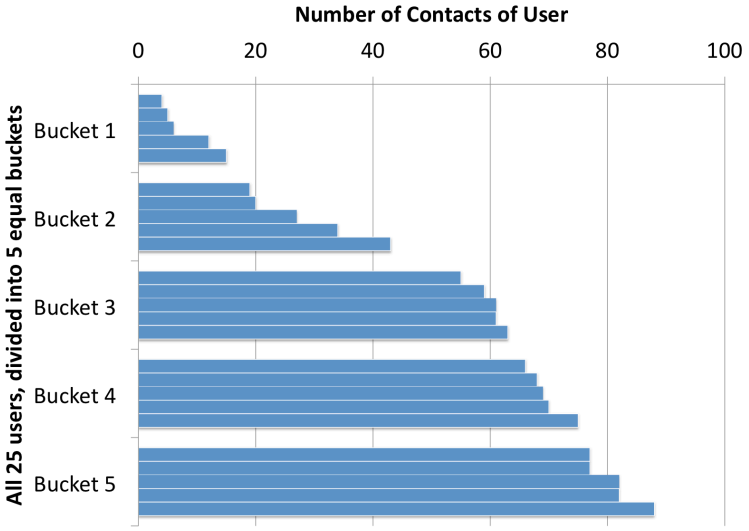
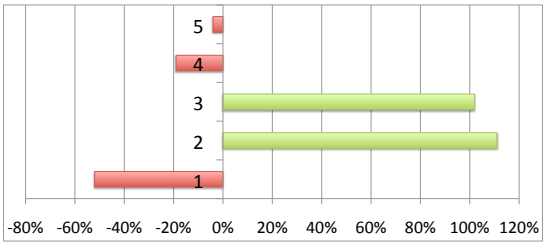


FIGURE 3.9: Example of user bucketing, based on a user's total social contacts. The 25 total users were split into 5 groups based on the number of contacts they had, with Bucket 1 containing users with fewest contacts and Bucket 5 with the most.

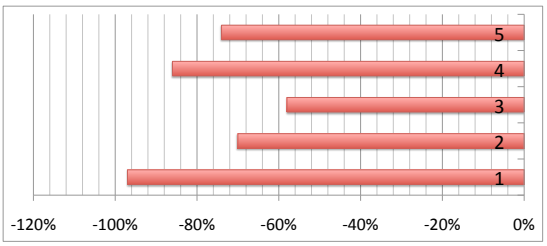
In order to observe the influence of the size of a user's available social context on performance, the users are divided into buckets based on their attributes and compare them. For the personal context, users are grouped based on how many photos they have; for the social contact context I divide into buckets based on the number of contacts the users have; and for the social group context I use the number of groups to which they belong.

Figure 3.11 shows the relative performance with respect to Mean Average Precision of the Personal Context, Social Contact Context, and Social Group Context compared to the Collective Context using topic sets where I partition the users based on photo count, contact count and group membership count. The 25 users are divided into 5 equally sized buckets based on increasing contact "count" as shown in Figure 3.9. The users are partitioned in the same way for each context.

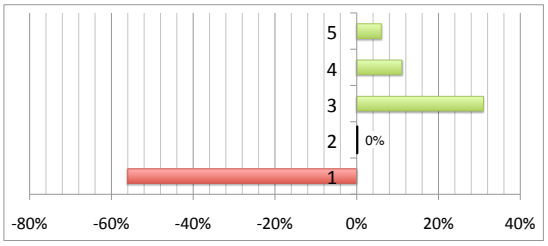
Figure 3.10(a) shows the performance of the Personal Context compared to the Collective Context for users with increasing number of photos. For users with relatively few photos, the Collective Context outperforms the Personal Context. However, for users with a medium number of photos (buckets 2 and 3) the Personal Context outperforms the Collective Context. For users with many photos the performance of the two contexts is similar.



(a) Relative performance of Personal Context compared to the Collective Context depending on the user's photo count bucket.



(b) Relative performance of Social Contact Context compared to the Collective Context depending on the user's contact count bucket.



(c) Relative performance of Social Group Context compared to the Collective Context depending on the user's group count bucket.

FIGURE 3.10: Evaluation of performance of the different contexts with respect to user attributes. The performance is measured in relative difference in MAP. Rows signify equally sized buckets of users, as shown in Figure 3.9

Figure 3.10(b) shows the performance of the Social Contact Context compared to the Collective Context for users with increasing number of contacts. The Social Context is poor for all groups.

Figure 3.10(c) shows the performance of the Social Group Context compared to the Collective Context for users with increasing number of group memberships. For users who are members of few groups the Social Group Context is clearly inferior than the Collective Context. However, for users who are members of a medium number of groups (bucket 3) the Social Group Context does improve over the Collective Context. For users who are members of many groups these two contexts perform similarly.

(a) Borda Count combination results

Run	**MRR**	Δ% over CC	Δ% over PC+CC	**P@5**	Δ% over CC	Δ% over PC+CC	**MAP**	Δ% over CC	Δ% over PC+CC
PC+CC	0.3445	29%‡	-	0.1856	78%‡	-	0.1049	63%‡	-
PC+SCC+CC	0.3365	26%‡	-2%	0.1672	61%‡	-10%†	0.0974	52%‡	-7%†
PC+SGC+CC	0.3893	46%‡	13%‡	0.1896	82%‡	2%	0.1140	78%‡	9%†
PC+SCC+SGC+CC	0.3866	45%‡	12%†	0.1840	77%‡	-1%	0.1109	72%‡	6%

(b) Multi Layered Perceptron combination results

Run	**MRR**	Δ% over CC	Δ% over PC+CC	**P@5**	Δ% over CC	Δ% over PC+CC	**MAP**	Δ% over CC	Δ% over PC+CC
PC+CC	0.3569	34%	-	0.1824	75%	-	0.1027	60%	-
PC+SCC+CC	0.3447	29%	-3%	0.1832	76%	<1%	0.0962	50%	-6%
PC+SGC+CC	0.3654	37%‡	2%	0.2008	93%‡	10%†	0.1093	70%‡	6%
PC+SCC+SGC+CC	0.3572	34%‡	<1%	0.1912	84%‡	5%	0.1008	57%‡	-2%

TABLE 3.3: Feasibility Study Combination Results
Performance metrics of combined context runs with (a) Borda Count (BC) and (b) Cost Weighted Multi-Layered Perceptron (MLP) with percentage change over Collective Context alone. MAP values marked with † are statistically significant with $p < 0.05$, and those with ‡ with $p < 0.01$

3.5.2.2 Performance of combined Contexts

Table 3.3 shows the results of combining various contexts using the different rank aggregation methods outlined in Section 3.3.7. For the sake of clarity, only the two best performing combination methods for this feasibility study with respect to my chosen metrics are shown—those which use the Borda Count and the Multi Layered Perceptron. The individual results are shown as well as their comparisons to two baselines: the Collective Context alone and the combination of the Personal and Collective Contexts for each method. The aim of this part of the evaluation is to investigate whether the Social Group Context and Social Contact Context can add to the performance of the system when used in combination with the Personal and Collective Contexts.

Combining the Personal Context and the Collective Context gives a statistically significant improvement beyond both the Collective Context baseline and Personal Context alone

when combined with the Borda Count method. When the Social Group Context is added, performance increases even further with respect to all our metrics, demonstrating that this particular type of social data can be useful in boosting performance over established evidence sources.

If, however, the Social Contact Context is combined with the Collective and Personal Contexts, we see a statistically significant degradation in performance for the Borda Count combination with p-value < 0.05, but an insignificant difference with the MLP combination with the same p-value. The Social Contact Context appears to perform so badly that it is in fact deleterious when used in combination with other contexts.

By combining all contexts together we see a (statistically significant) increase in performance over the individual contexts alone, but overall performance is still marginally lower than the Personal+Social Group+Collective combination, most likely because of the inclusion of the harmful Social Contacts Context.

3.5.3 Feasibility study evaluation

By performing this initial study I have been able to:

- gain insight into the scale and topological structure of the networks involved in the four different contexts
- demonstrate the feasibility of computing the data required to produce recommendations using my proposed probabilistic framework
- produce early indications of how the variation in user attributes such as the number of contacts, groups and uploaded photos effect tag recommendation performance with respect to each context
- select a combination method (Borda Count) that performs well across my three chosen evaluation metrics

While this study showed that the system is capable of making better suggestions than established baseline equivalent systems, 25 users are too few to draw any more significant conclusions as to general performance of this framework. Also, the approximation of my probabilistic framework used in this study, while only slightly different from that used in the full experiment, is likely to have resulted in reduced performance.

In light of these findings, a more comprehensive experiment can be justified that looks at a much larger set of users with more diverse attributes, using the Borda Count method to combine individual context runs and that, if possible, doesn't have to compromise during the probability calculation stages.

3.6 Second stage: experiment

3.6.1 Data collection

The source collection for this experiment is comprised of the annotations of over 700 million public Flickr photos, uploaded before May 2008. 300 hundreds users were selected for evaluation. This number was chosen as it was significantly larger than in the feasibility study, providing more data to support any final conclusions, particularly with respect to the buckets of users of varying attribute values. The users were select at random from among users that represented a variety in "socialness" as per the initial study. The users are divided into buckets based on how many contacts they had:

Bucket 0: Users with zero contacts.

Bucket 1: Users with 1 or 2 contacts.

Bucket 2: Users with 3 to 10 contacts.

Bucket 3: Users with 11 to 50 contacts.

Bucket 4: Users with 51 to 250 contacts.

Table 3.4: Statistics over the 300 users in our experiment.

Statistic	Min.	Max.	Mean	StDev
Number of Contacts	0	1472	122.7	243.7
Number of Groups	0	656	89.8	135.5
Number of Photos	102	94415	1185.9	5586.9

Bucket 5: Users with 251 contacts or more.

From each bucket I selected 50 users who satisfy the following criteria:

- They have at least 100 photos in the data collection. While this focuses our evaluation on active users, it also provides more data to analyse and therefore is more likely to provide a firm foundation for any resultant conclusions.
- They have at least 20 photos uploaded after May 2008 that satisfy the following criteria: 1) the photos need to have at least 10 tags; 2) no two photos have the same tag-set. From the resulting photos, 10 were randomly chosen for testing.

The evaluation collection therefore contains 3,000 photos from 300 different users. Table 3.4 shows some characteristics of the 300 users in terms of the number of contacts they have, the number of groups to which they belong and the number of photos they have in the dataset.

With so many more users than in the feasibility study, the amount of data to be processed increased considerably. Fortunately, due to the advances in data processing techniques outlined in Section 3.3.8, the increase in data processing was not a problem when the time- and space-efficient PIG scripting environment was used. More significantly, the change in data processing also meant that the compromise in the probabilistic framework used in the initial study was no longer required and all possible candidate tags could be exhaustively evaluated.

The processing was done on the same Hadoop cluster of 100 nodes with processing taking from a few hours to a day, depending on the context that was being computed, with the Personal Context being quickest, and the Social Group Context being the slowest.

The experiment design was the same as that for the feasibility study except for the expanded dataset used, the improvement in the probability calculations and the more efficient mechanism used for calculation. Again the tag recommendation system is evaluated on a set of photos uploaded after May 2008 to ensure that there was no overlap between the set of photos for which our co-occurrence statistics are calculated and the set of photos used for the evaluation. As for the feasibility study, the evaluation collection for this experiment was created using the publicly accessible Flickr API[16].

3.6.2 Results

3.6.2.1 Performance of individual Contexts

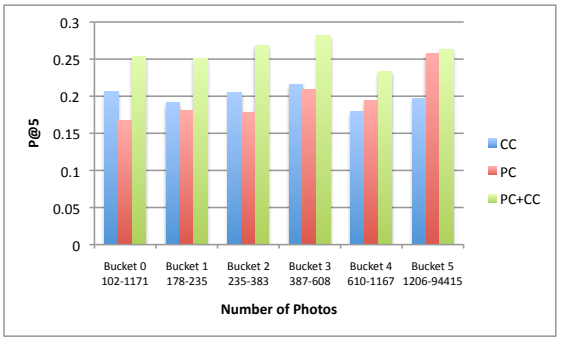
Table 3.5: Evaluation results for the individual contexts. Improvement is calculated relative to the Collective Context baseline. Values marked with †are significant with $p < 0.05$ and those with ‡with $p < 0.01$.

Run	MRR		P@5		MAP	
Collective Context	0.4473	–	0.1991	–	0.0934	–
Personal Context	0.3459	-22.7% ‡	0.1979	-0.6%	0.1034	10.7% ‡
Social Contacts Context	0.0997	-77.7% ‡	0.0413	-79.3% ‡	0.0171	-81.7% ‡
Social Groups Context	0.3395	-24.1% ‡	0.1585	-20.4% ‡	0.0777	-16.8% †

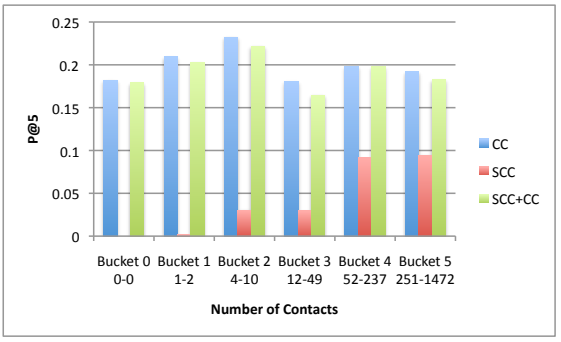
The results of evaluating different contexts in isolation are shown in Table 3.5. It can be seen that, when measured over all users and their queries, the personalised contexts mostly perform significantly worse than the non-personalised Collective Context. The Social Contacts Context is particularly bad when considered on its own. The MAP of our Personal Context run is, however, significantly higher than for the Collective Context. This implies that, on average, the relevant tags suggested by the system in the Personal Context occur higher up the ranked list of returned results than for the Collective Context. This is a particularly valuable finding as early precision is important within the use-case of this experiment.

[16] http://www.flickr.com/services/api/

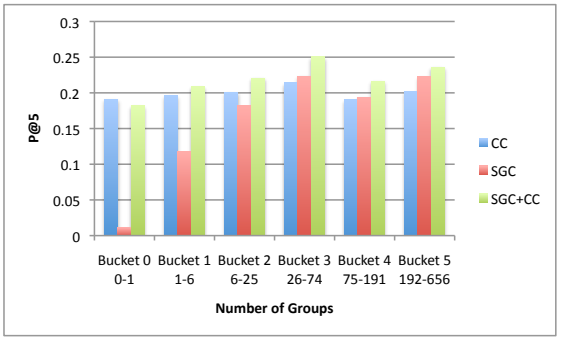
The results in Table 3.5 don't describe the relative proficiencies of the contexts for individuals—this invites the question: "Are some contexts better than others for certain types of users?" To explore this, I extend the analysis by looking at sub-sets of users based on social criteria.



(a) Relative performance of Personal Context (PC) compared to the Collective Context (CC) depending on the user's photo count



(b) Relative performance of Social Contact Context (SCC) compared to the Collective Context (CC) depending on the user's contact count



(c) Relative performance of Social Group Context (SGC) compared to the Collective Context (CC) depending on the user's group count

FIGURE 3.11: Evaluation of performance of different contexts depending on the user characteristics. The performance is measured in terms of P@5. Columns signify equally sized buckets where each bucket contains 50 users. The bucket ranges are also shown.

With respect to our user interaction scenario, higher priority should be given to early precision than for recall, as explained in Section 3.4.3. Unlike in the feasibility study where I used MAP as an indicator of general performance, in the following analysis I focus on the performance metric of 'Precision at 5' that more closely matches the requirements of the use-case scenario of the experiment.

Figure 3.11 shows the relative performance with respect to P@5 of the Personal Context, Social Contact Context, and Social Group Context compared to the Collective Context and

their combination with the Collective Context. Using sets partitioned on the users based on photo count, contact count and group membership count, the 300 users are divided into 6 equally sized buckets based on increasing "count" (it must be noted however that the particular users in each bucket vary between each graph).

Figure 3.11(a) shows the performance of the Personal Context compared to the Collective Context and their combination, for users with increasing number of photos. Bucket 0 contains the 50 users with fewest photos and bucket 5 contains the 50 users with the greatest number of photos. It can be seen that for users with relatively few photos the Collective Context outperforms the Personal Context. However, for users with many photos (buckets 4 and 5) the Personal Context outperforms the Collective Context. This suggests that a user's personal tag dictionary, whilst personalised, does not become more useful for tag recommendation than collective knowledge until it reaches some critical size (suggested by the graph to be derived from between 387-1167 photos). From then on it is sufficiently large and well tailored to the vocabulary of the given user and is capable of providing better tag recommendations.

Figure 3.11(b) shows the performance of the Social Contact Context compared to the Collective Context and their combination for users with increasing number of contacts (i.e., bucket 0 contains the users with the fewest number of contacts and bucket 5 contains users with the greatest number of contacts). The Social Context is poor for all groups and always detrimentally affects the combination run. As was indicated in the feasibility study (see Section 3.5.2.2), the tagging behaviour of a user's contacts seems to poorly reflect that of the user and so is unhelpful when making tag recommendations.

Figure 3.11(c) shows the performance of the Social Group Context compared to the Collective Context and their combination for users with increasing number of group memberships (i.e., bucket 0 contains the users who are members of the fewest groups and bucket 5 contains the users who are members of the largest number of groups). For users who are members of few groups the Social Group Context is clearly inferior to the Collective Context.

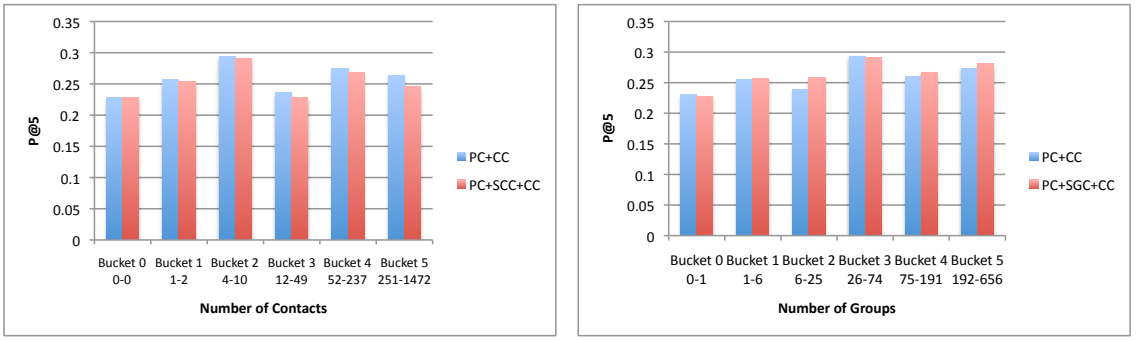
TABLE 3.6: Evaluation results for the combined contexts. Improvement is calculated relative to the PC + CC baseline. Values marked with †are significant with $p < 0.05$ and those with ‡with $p < 0.01$.

Run	MRR		P@5		MAP	
PC+CC	0.5307	–	0.2587	–	0.1347	–
PC+SCC+CC	0.5189	-2.2%‡	0.2527	-2.4%‡	0.1300	-3.5%‡
PC+SGC+CC	0.5406	1.9%‡	0.2638	2.0%†	0.1351	0.3%
PC+SCC+SGC+CC	0.5260	0.9%	0.2591	0.2%	0.1319	-2.1%†

However, as group membership increases, performance tends to increase. For users who are members of many groups (buckets 3 – 5) the Social Group Context does improve over the Collective Context. This suggests that with a sufficiently large collection of groups from which to mine tags (suggested by the graph to be between 6 and 74 groups), useful recommendations can be made. It also seems to lend support to the intuition that groups are likely to reflect the interests of a user, that ultimately affect or reflect their tagging behaviour.

Similar trends as described above are reported by Konstas et al. (2009) where, in the context of music recommendation, the music taste of one's friends is less likely to positively correlate with their music taste. Conversely it is possible to make good recommendations based on other users that share the same taste.

3.6.2.2 Performance of combined Contexts



(a) Relative performance of Personal + Collective Contexts combination and Personal + Social Contacts + Collective Contexts combination depending on the user's contact count

(b) Relative performance of Personal + Collective Contexts combination and Personal + Social Group + Collective Contexts combination depending on the user's group count

FIGURE 3.12: Evaluation of performance of different contexts depending on the user characteristics. The performance is measured in terms of P@5. Columns signify equally sized buckets where each bucket contains 50 users. The bucket ranges are also shown.

Table 3.6 shows the results of combining various contexts using the Borda Count method outlined in Section 3.3.7, chosen due to its performance in the feasibility study. The individual results are shown as well as their comparison to a baseline of the combination of the Personal and Collective Context. The aim of this part of our evaluation is to investigate whether the Social Group Context and Social Contact Context can add to the performance of the system when used in combination with the more conventional Personal and Collective Contexts.

The combination of the Personal Context and the Collective Context gives a highly performing baseline with which to compare the other runs. Referring back to the example scenario illustrated in Section 3.4, a P@5 of 25% implies being able to exactly match 2 tags in a prediction set of size 8.

If the Social Contact Context is combined with the Collective and Personal Contexts, a statistically significant degradation in performance is seen for the combined run with p-value < 0.01 for all metrics. The Social Contact Context appears to perform so badly that it is harmful to overall performance when used in combination with other contexts. This further supports the findings in the previous section that the tagging behaviour of contacts is unhelpful for making tag suggestions, as well as validating the results from the feasibility study.

When the Social Group Context is combined with the Personal and Collective Contexts a marginal improvement can be observed, but only statistically significant for MRR and P@5. This suggests that there is some value in using the Social Group Context for tag suggestion.

By combining all contexts together a statistically insignificant change in performance is seen over the combined baseline for MRR and P@5, and a significant decrease in MAP. The inclusion of the harmful Social Contacts Contexts would explain the decrease in performance when compared to the Personal, Social Groups and Collective combination.

3.7 Conclusions

In this chapter I have demonstrated how personal tag co-occurrence data can be used to provide more relevant recommendations of tags to a user when annotating photos than the baseline systems found in related work. In doing so, along with Drs. van Zwol and Sigurbjörnsson, I have formalised a flexible, extendible model that describes user interaction with photos and other users as a set of graphs and characterises the tag usage of different social aspects of an online community like Flickr. I have further shown that by combining the personalised graphs with data from all users of Flickr, I can significantly improve performance of tag suggestion when compared to existing state of the art techniques.

I have provided evidence that addresses the second of the sub-questions of my hypothesis regarding identifying which social connections are most valuable to recommender systems. I have shown that while intuition suggests that the interests of a user's contacts may reflect their own, the Social Contacts Context as defined here has very little value when it comes to making tag suggestions, by itself or in combination with other contexts. In addition, and most interestingly, I have demonstrated the considerable usefulness of additional social contextual data, in this case the Social Group context.

With respect to my third hypothesis sub-question regarding the effective use of social interaction data, I have presented a framework for extracting tag co-occurrence graphs from different 'strata' of a user's social graph from Flickr and shown how this can be evaluated with respect to established information retrieval performance measures. The framework can be extended with additional contexts to gain a better understanding of the relative usefulness of social graphs defined by other inter-user relationships not investigated here, perhaps including a wider range of both direct and indirect relationships like shared commenting activity and mutual Favourite image labelling.

The model I have presented has multiple benefits over non-personalised, non-social aware systems, including some that are less immediately obvious. For example, users who do not

use English while interacting with Flickr benefit from a system that focuses on their past tagging behaviour. I am able to make relevant recommendations in their own language by virtue of their past interactions that make up their personal tag set and the interactions of their social groups, in addition to the most popular (usually English) tags contributed by the generalised data.

This experiment has also highlighted the difficulty in selecting social data from Flickr that are ultimately useful when trying to boost performance for this particular user task. I am confident that through further exploration of the rich social data available within online media sharing sites like Flickr, I could improve performance further still. Learning weightings for the combination of our different contexts could be done on a more sophisticated, per user level which could also increase the ability to make good tag recommendations—an area that could be investigated in future.

3.8 Reflection on tag suggestion experiments

This chapter has shown that large web-scale data can be effectively mined for information that can be used to support the user when annotating their media by suggesting the kinds of tags they would add based on their past behaviour. These tag recommendations in turn help improve the quantity and quality of annotations within Flickr, thereby making such a system easier to navigate by users and easier to manage by the service providers. By making it easier to extend the annotation of images, the number of photos without any tags or with only few that currently exist in such systems could be made accessible once more, increasing their value to the community.

The problem of encouraging users to annotate their media is common to many online media sharing systems, particularly those that encourage mass uploading of photos. The conclusions drawn in this chapter transfer to systems other than Flickr that share the same kind

of data and user interaction. For those systems that are based on extensive and comprehensive social networks like Facebook, the performance of my approach could be significantly higher compared to existing tagging mechanisms due to the greater availability of rich social interaction information.

I have shown how the various tag networks perform, relative to each other. In particular, the relationship between the two personalised social contexts, the SCC and SGC, has demonstrated that some interpersonal relationships are more valuable than others when it comes to tag suggestion. The SCC is, in essence, a reflection of a real-world social network that has been reapplied online, whereas the SGC is created anew when they start interacting with others in Flickr. This may indicate a general trend for this type of task, in that the relationships people form by interacting with others inside the system are more valuable than those that have merely been inherited from outside of it.

This would suggest that the value of a social network is dependent on not just the system it is developed in but also on the interaction on which it is based. This has implications for online services that gather users based on their interactions in other systems (e.g. importing friends from a social profile directory like Facebook) and try to extract value from them in a different context (e.g. disseminate information through links via Twitter).

This could be generalised further to non-online, non-computerised environments. As an example, imagine an organisation like a university with a large body of academic research staff. These staff members have different skills, expertise and areas of interest with respect to their research projects. They work together, socialise together and generally interact in a number of ways, but may not necessarily be aware of all the other people who share their attributes. The model proposed in this Chapter could be applied to such an environment, so that where Flickr has users who share photos that are described by tags, a university research would have projects that they work on with others, and these projects have particular themes or focuses. Once the model is applied, the resultant graphs could be used to identify clusters of people with shared research interests and would thus provide the organisation

with information that could be used to connect previously disconnected people and teams, potentially increasing productivity.

Ultimately, by treating the combined interactions of users within an online community such as that in Flickr as a multi-layered social multigraph, it is possible to tease out trends and patterns that can ultimately lead to insight into tagging behaviour. In the case of the experiments in this chapter, these trends can help inform systems that focus on finding media according to its semantic content, once a particular information need or query has been formed. However, this ignores other interaction types like browsing, as well as the more complex nature of the visual media being stored and indexed. By understanding more about what users find 'attractive' in an image, other interactions can be supported and improved. By taking advantage of the visual and social contextual information that can be derived from images, in addition to traditional textual annotation, it is possible to do this in such a way that out-performs existing systems.

This is theme of the following series of experiments found in the next chapter.

Chapter 4

Identifying Flickr Favourites Using Social Context

> *"Nah," he said, eventually. "I've looked at the colours on flowers. They're definitely built-in."*
>
> (Terry Pratchett, Diggers)

Roadmap In this chapter I introduce the field of image recommendation and propose a supervised machine learnt approach that is trained using the Favourite label in Flickr. I extract a range of social, textual and visual features from the data of 400 users to produce a model that accurately identifies Favourite-labelled images, using two alternative approaches for handling the lack of negative feedback available in Flickr.

The value of individual features is analysed and I show how social features come to dominate my classification model. I evaluate both a single classifier for all users, but also train classifiers on an individual basis, and show when and why the use of these two approaches is most appropriate.

Note: This chapter is based on work that was predominantly carried out while working in the laboratory of Yahoo! Research Barcelona with Dr Roelof van Zwol and Lluís Garcia Pueyo. As such, some sections of this chapter closely reflect the content of the papers that were published on this work, particularly the *ACM Multimedia 2010* paper (van Zwol et al., 2010).

4.1 Introduction

As of May 2007, Flickr had over 2 million new photos uploaded every day from around 8.5 million registered users and served out 12,000 photo per second during peak times (van Zwol, 2007). The number of users has since increased to over 32m (Flickr, 2009), almost a four-fold increase in two years. By October 2009 Flickr comprised of over 5 billion uploaded images (Sheppard, 2010).

With such a large collection of images, finding those that best match the information needs of users (from specific, explicit queries to supporting browsing for enjoyment) becomes more difficult. Not only are the datasets themselves growing, but the ways users want to interact with them continues to expand. For example, new ways of interacting with this data have arisen, particular in the area of mobile devices, but also by making data available programmatically through public APIs, as in the case of Flickr[1] and Picasa Web Albums[2]. This enables and encourages the development of new third-party user interfaces, data managers and image processors of which the original system designers may not have conceived. The personalised tag suggestion framework demonstrated in the previous chapter is an example of such an added-value service built on publicly available data.

In addition to the changing uses of these online image collections, the users themselves are diversifying. The early adopters of many new technological systems tend to come from

[1] http://www.flickr.com/services/api/
[2] http://http://code.google.com/apis/picasaweb/

a relatively small demographic and, as time has gone on and systems have become more popular, a wider range of users has become part of the community.

At the time of writing, many systems still tend to use the same paradigm of providing services to users in the same way they did when they had smaller datasets being used by a small, more homogenous community of people using limited ways of interacting with the system. They do not take into account and fully exploit the growing variation between users, nor do they take advantage of the affiliative relationships that arise within these burgeoning communities.

This situation raises the following questions:

- Can we learn what individual users find interesting, that they like or would pick out from the flood of possible data as something particularly special or relevant?
- Can the mental burden on the user be reduced by tailoring their results to their specific needs?
- To what degree is social context useful in this personalisation?

In the previous chapter, I was able to demonstrate the value of social context as a way of understanding user behaviour and how it can be used to augment and improve existing systems for supporting users. In particular, I looked at two social relations: 'contact' and 'shared group membership'. This chapter takes that theme further by comprehensively investigating many different types of social connections and their value in learning user image preference. As image preference is a highly user-specific phenomenon influenced by many factors, I use a personalised learning approach.

The analysis from the tag suggestion experiment also showed how complex the interaction of different features can be and this chapter will look closely at the combination of feature types, namely social, visual and textual. By doing so, I will provide supporting evidence to help answer the third hypothesis sub-question as defined in Section 1.2:

How can different kinds (textual/visual/social) of media/user descriptors be combined effectively in a image recommender system?

As for the tag suggestion experiment, I use an existing user support paradigm—in this case a recommender system—to allow my work to be comparable to those of others, as well as to assess my findings in a way that is directly related to the problems found in web-scale media sharing environments.

The Flickr Favourite label

In many online photo sharing systems, users post photos in various shared interest groups, tag other people's photos, provide ratings and give comments on photos they like. In addition, users can mark a photo with a positive feedback label like the Flickr Favourite[3] label, the Facebook "Like" button or the Picasa Web "⋆" (star) button.

The use of this kind of label is relatively rare when compared to the scale of the datasets involved, as the labels are only assigned when the users consciously decide to add them to media, and not all users are aware of this particular labelling facility, or choose to use it. This type of single positive feedback label is one of the few explicit ways—at the time of writing at least—that users can give information back to the system about photos they like, want to bookmark or feel some connection with. While this kind of label may be vague, and indeed the sites with such labels seldom formally prescribe how exactly they should be used, they do provide a rare signal that indicates what users want from their interactions with these photo collections.

The concept of relevance is explored in Chapter 2 where I show that there is no single definition used throughout the community and also how any definitions that do exist are dependant on the context of the information need in question.

[3]Throughout this Chapter the use of the word *favourite* will refer to the specific binary "Favourite" annotation used in Flickr, unless otherwise noted.

I propose the Flickr Favourite label to be a proxy for user preference within the confines of the Flickr environment. If it can be shown that this is the case, a system which can predict favourite images would help users find relevant images. If it can be shown that the Favourite label can be used to find images of relevance for Flickr users in particular, it is plausible that the label's analogues in other similar systems could do the same. This will also provide evidence to answer my fourth hypothesis sub-question.

> *Can single positive feedback cues like the Flickr Favourite label be used to train systems to predict further Favourites?*

While the Favourite label (and its analogies in other systems) is perhaps merely an indicator of a user's connection to the photo, it does not elaborate on which attributes of the photo make them want to label it so. To learn what it is about an image that makes a user label it a Favourite, the image must be described and features selected that can be used to learn what influences their decision, a task that my machine learnt approach addresses. By analysing the value of the catalogue of visual, textual and in particular social features I have selected and employed, I will provide supporting data to help answer the first of my hypothesis sub-questions.

> *Which social connections yield the most valuable information for use in tag and image recommender systems designed for large online photo sharing systems?*

4.1.1 Problem specification

There are many reasons why a Flickr user may label a photo a Favourite. These motivations include feeling a resonance with the semantic content or description of the image, e.g., the user likes wildlife, buildings, or landscape photography. Images are inherently visual media and so the aesthetics of a photo will also be important. I propose also that the existence and type of social connection between the photo owner and the user that marks the photo as a Favourite will affect their decision as well.

Whereas existing work has attempted to use textual and visual information in recommender systems it has mostly ignored the social context of the user within the system and their interactions with others. It is the introduction and evaluation of social features in combination with both textual and visual features that I show in this chapter to have particular value when tackling this problem.

After extracting these features for a set of users that have marked photos as Favourites, I train a machine learnt classifier to be able to identify these images. I use images that have been labelled as Favourites as positive examples, but as there is no equivalent label for non-Favourite images from users in Flickr, I propose and generate two plausible scenarios to address this.

For one scenario, I collect the photos labelled as Favourites from a set of 400 users of varying Flickr activity, and I add randomly selected, non-Favourite photos from throughout Flickr to provide the negative examples.

The other scenario contains the same positive examples for each of our users, but the non-Favourite photos are selected at random from the social network of each user. I include the second scenario as users are generally more likely to be exposed to photos of their contacts and from the groups in which they participate than completely random photos, due to user behaviour and the interface design of Flickr, as highlighted by the work of Lerman and Jones (2006) and Lerman (2006).

Through an empirical evaluation I measure the effectiveness of my system's predictions in terms of selected standard metrics. I compare the performance of the runs across both scenarios, for users of all activity levels as well as compare the difference in performance between training a single, small classifier generalised for all users and individually trained classifiers for each user.

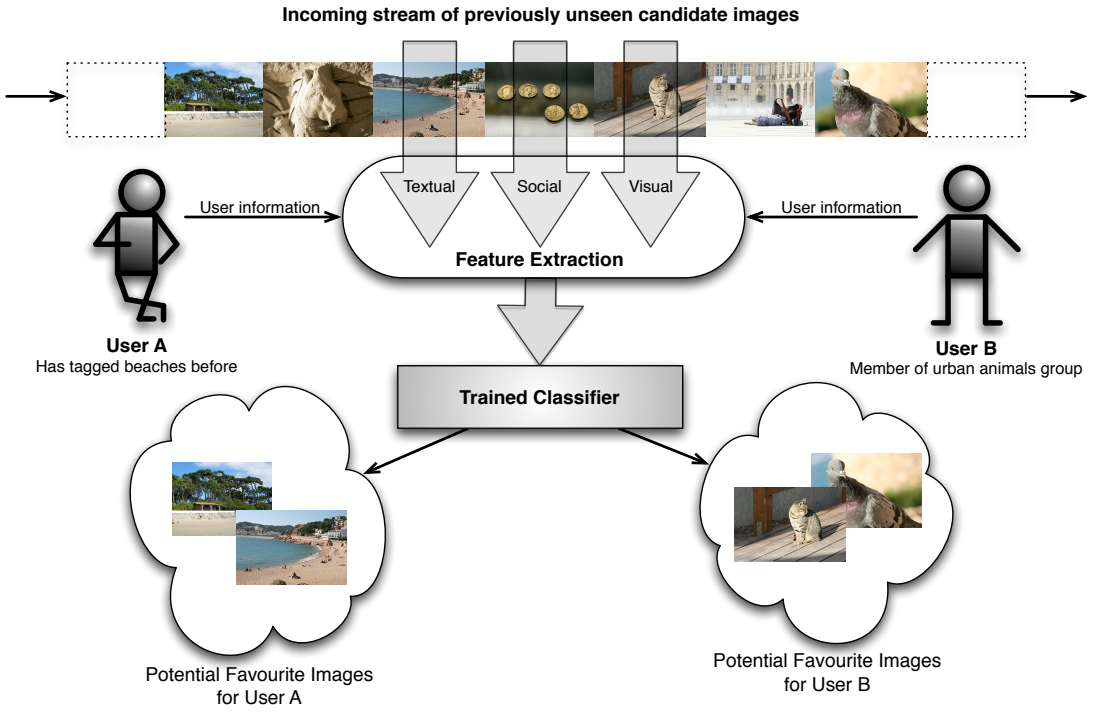


FIGURE 4.1: Work flow for predicting Favourite photos.

4.1.2 Predicting Favourite photos

I envisage a scenario where a user is exposed to an incoming stream of photos that are being uploaded into a system like Flickr. Obviously this would include many diverse images, few of which would interest the user and so I use my trained classifier to make a judgement for each incoming photo based on the features extracted from the image. Those that are judged as likely to be relevant are then show to the user. This might be in isolation, or these photos may be emphasised among the others while remaining part of an incoming photo stream. A diagrammatic overview of the approach I adopt is given in Figure 4.1.

4.2 Experiment

4.2.1 The multi-modal feature space

I propose a multi-modal[4] approach that uses the textual, visual and social signals that I consider to play important roles in a user's annotation decisions.

Based on a randomly sampled inspection of the Favourite labelled photos in an available dump of 250 million recent (May 2008) Flickr photos that had two or more tags[5], I hypothesise that:

1. Favourite images tend to be visually pleasing, e.g., sharp, large, vibrant, etc., but not consistently so.
2. Many users focus on a small number of topics of their interest, like children, cityscapes, flowers, nature, or portraits.
3. In many occasions Favourite photos are posted in a group to which the user is subscribed, or the photo is taken by one of the contacts in that user's social network.

These observations correspond with the three feature classes used in my machine learnt classifier.

4.2.2 Supervised learning for classification

I treat this experiment as a classification task—being able to distinguish Favourite from non-Favourite images. Initially a single model is trained for all users. This approach is scalable in that only one model needs to be trained regardless of the number of users who use it. A single classifier approach is also capable of assisting users that have not actively

[4] I use the term **multi-modal** to describe a system that uses features that describe multiple *aspects* of an image—visual, semantic, social, etc.—as opposed to the definition occasionally used elsewhere that describes systems that handle multiple *channels* of media interaction— visual, audio, tactile, etc.

[5] This is the same dataset that was made available and used in the previous chapter, hence the tag criteria.

been labelling photos as their Favourites (the majority of users), as it is not dependent on the behaviour of specific users, but on trends found among the many users on which it was trained.

In contrast, I also train individual models for each user in order to see whether the specificity of such trees provides any improvement over the general classifier.

In order to produce an effective classifier, a classification mechanism must be chosen appropriate to this task and the data being used. In the literature related to machine learning, there are many reviews comparing classes of classification techniques suitable for large real-world problems using test datasets and evaluating results with respect to a range of metrics. Each attempts to give guidance as to which technique should be chosen given the statistical characteristics of the data involved and the task criteria.

An early example of such a survey is the StatLog project in which King et al. (1995) evaluated symbolic learning, statistical and neural network based algorithms. In addition to evaluating each of the chosen algorithms with each of their experimental datasets, the authors highlight a few key findings. They show that the 'best' algorithm for a given experiment is highly dependant on the nature of the data being used, in that they showed that accuracy varied significantly for a given algorithm when applied to different datasets. They also showed that in general the overall accuracy of the different algorithms did not vary significantly when compared to each other.

They suggested that datasets that exhibited high skewness (greater than 1) and had over 38% binary/categorical attributes would likely favour symbolic learning (the class that includes decision trees and their variants, amongst others).

The survey of King et al. (1995) was extended by Lim et al. (2000) who also look at additional performance relating to machine learning classifier algorithms such as training time and include a wider range of decision tree implementations as well as some new spline-based statistical approaches They also looked at adding independent noise to see how this effects

resultant accuracy as well as analysing scalability. They found that the differences in mean square error rates did not vary significantly between their highest performing algorithm and the lowest. They suggest that instead of focusing on overall accuracy when choosing a classifier algorithm, as this is not particularly discriminating, researchers should choose based on speed and memory usage. In addition, they highlight the additional value of decision trees with respect to interpretability—valuable when trying to understand how a trained decision tree reached a judgement.

Eklund and Hoang (2002) undertook another evaluation of the same three three classes of algorithms and found that the Linear Machine Decision Tree (LMDT) algorithm had the highest accuracy compared to the others they evaluated. The LMDT algorithm differs from simple trees like C4.5 (Quinlan, 1993) in that it trains a linear machine which then serves as a multivariate test for the decision nodes in the tree.

Caruana and Niculescu-Mizil (2006) undertook another survey, also evaluating old and newer algorithms, as well as extending their analyse to include the calibration of training algorithms using Platt's method for logistic regression (Platt, 1999) as well as isotonic regression They showed that such calibration could significantly improve performance with respect to their metrics. While some of the evaluated algorithms performed better on average than others, there was significant variability between datasets and metrics. Calibrated boosted trees were again shown to be the best performing algorithm.

A more recent survey was undertaken by Kotsiantis (2007), in which they make the succinct point which I quote here:

> *"The key question when dealing with ML [machine learnt] classification is not whether a learning algorithm is superior to others, but under which conditions a particular method can significantly outperform others on a given application problem."*

They also adopt the position that accuracy has ceased to be the primary decision criterion for choosing a classifier algorithm (their evaluation found their algorithms did not vary significantly in terms of accuracy), instead performance in terms of computational expense and training time become more pertinent.

Throughout all these surveys, from among the algorithms chosen by the authors, decision trees and boosted decision trees in particular were shown to do well more consistently. Based on the findings of these surveys, the most appropriate algorithm for the experimentation in this chapter is one that is quick to train, is space efficient and allows for easy interpretability and hence analysis. With these criteria in mind, I chose stochastic gradient boosted decision trees (GBDT) introduced by Friedman (2001) as my learning algorithm. The trees in stochastic GBDT are trained on randomly selected subsets of the training data and are usually less prone to over-fitting (Friedman, 2001) than older decision tree algorithms. As part of the boosting process, different loss functions can be used, and for the research presented here, least squares regression is used.

As mentioned before, decision trees lend themselves to being easy interpretable with respect to how they make judgements. In the case of GBDT, the relative importance of a given feature with respect to the other features can be quantified. Using the notation of Friedman (2001), in a single tree T, with L terminal leaf nodes, in the set of trees $\{T_m\}_1^M$ (that would form a complete ensemble GBDT classifier), the squared relative importance $\hat{J}_i^2$ of feature j is calculated as:

$$\hat{J}_j^2(T) = \sum_{t=1}^{L-1} \hat{I}_t^2 I(v_t = j) \tag{4.1}$$

where $\hat{I}_t^2$ is the improvement in squared-error due to the split at node t and $v(t)$ is the feature used for splitting at that node.

In other words, the squared relative importance of a particular feature is given by the sum of the squared improvements in squared-error it makes at each node in the tree in which it

is the decision variable.

For a the set of trees $\{T_m\}_1^M$, the overall importance of a feature $\hat{J}_j^2$ is then the mean average over all M trees:

$$\hat{J}_j^2 = \frac{1}{M} \sum_{m=1}^{M} \hat{J}_j^2(T_m) \tag{4.2}$$

As the importance of each feature is relative to the others, they are expressed as a value normalised to within the range of 0 to 100, where the most influential feature is given the value of 100 and the others are scaled accordingly.

The principle objective of the experimentation in this chapter is not to obtain the highest possible performance with respect to my metrics, but to measure relative impact of the different features I use for training, in particular the social features. For this reason, I do not attempt to optimise training parameters and fix them throughout, so that the maximum number of sub-trees is 300, maximum leaf nodes per tree is 30, the learning rate is set to 0.04 and the sampling rate is 0.5. For specific details of these parameters, I refer the reader to the paper of Friedman (2001).

4.2.3 Evaluation considerations

Individual binary classifier judgements for examples that are either positive (P) or negative (N) can be classed as being of one of four types, also known as the quadrants of a *confusion matrix*:

True Positives (TP) The number of correctly judged positive examples

False Positives (FP) The number of incorrectly judged positive examples (as known as the Type I error)

True Negatives (TN) The number of correctly judged negative examples

False Negatives (FN) The number of incorrectly judged negative examples (as known as the Type II error)

The combination and proportions of these four values form the basis for the standard binary classification metrics of:

Sensitivity $\frac{TP}{TP+FN}$, describes the classifier's ability to identify positive results. High sensitivity implies a low Type II error rate. It is analogous to *recall* in information retrieval

Specificity $\frac{TN}{TN+FP}$, describes the classifier's ability to identify negative results. High specificity implies a low Type I error rate.

Positive Predictive Value (PPV) The proportion of positive examples correctly identified as such. In information retrieval, PPV is analogous to *precision*.

Negative Predictive Value (NPV) The proportion of negative examples correctly identified as such.

Accuracy $\frac{TP+TN}{P+N}$, The proportion of correctly judged examples in the whole population.

Receiver Operating Characteristic (ROC) curve is also a common way of interpreting true positive rate (sensitivity) vs. false positive rate while the classification discrimination threshold is varied. The shape of the plotted curve for a good classifier should favour the [high sensitivity, low (1-specificity)] corner of the plot with respect to the positive diagonal. ROC curves are not used to give an indication of overall performance, but to show how performance varies according to some continuous random variable such as the discrimination threshold in a binary classifier, which in turn can be used to select values for such a variable.

With respect to the classifiers trained in the experiments in this chapter, the metrics must be appropriate to the task they are evaluating.

For example, the task could be considered a simple binary classification task, in which each judgement could be considered individually and therefore metrics like sensitivity and specificity would be appropriate.

However, I envisage that the image recommendations made by such a classifier would be used to show the user multiple images at once. For example, a user would be shown a page of multiple images that have been selected as likely to be labelled as Favourites for that user.

Therefore, the task can also be considered to be analogous to a retrieval task, for which metrics like precision and recall are appropriate. Within this context, the objective of this experiment becomes the optimisation of precision of the set of photos judged to Favourites, while maintaining an acceptable level of recall (ensuring as many potential Favourites are correctly judged). In addition, due to the choice of using binary gradient boosted decision trees as the training algorithm, there is no discrimination threshold to alter and so ROC curves would not be appropriate.

As a way of summarising both precision and recall I also use:

Average F-measure The average F-measure, F_{avg}, reports the weighted harmonic mean for the precision/recall over both the positive and negative classes, treating them equally, computed as follows:

$$F_{avg} = R_{neg} \times \left(\frac{(1+\beta^2) \times p_- \times r_-}{\beta^2 \times p_- + r_-} \right) + R_{pos} \times \left(\frac{(1+\beta^2) \times p_+ \times r_+}{\beta^2 \times p_+ + r_+} \right)$$

where β represents the ratio of the importance of recall to that of precision (1 when equal, 0.5 when precision is twice as important as recall, etc.), as defined in the book of Van Rijsbergen (1979),"Information Retrieval". I use $\beta = 0.5$ to reflect the relative importance of precision compared to recall. R_{pos} and R_{neg} are variables with positive real values between 0 and 1 such that $R_{pos} + R_{neg} = 1$, that allow the F_{avg} metric to be tuned to focus on one class over the other.

Trivial classifier for comparison In order to give context to the output of the classifiers trained in this chapter, I propose the following trivial classifier against which my results

will be compared.

This trivial binary classifier randomly assigns either the positive or negative class with a ratio of positive to negative instances of $p : n$. Using the ratio used in the experiment of $1 : 6.5$[6], the expected probabilities become:

- P(positive) $= \frac{p}{p+n} = \frac{1}{7.5} = 0.1\dot{3}$
- P(negative) $= \frac{n}{p+n} = \frac{6.5}{7.5} = 0.8\dot{6}$

Assuming random class assignment with respect to these probabilities, and assuming the same proportion of Favourite and non-Favourites used in the experiment (1:6.5), expected precision, recall and $F0.5_{avg}$ values can be precomputed:

	Favourite	Non-Favourite
Precision	$0.1\dot{3}$	$0.8\dot{6}$
Recall	$0.1\dot{3}$	$0.8\dot{6}$
$F0.5_{avg}$	0.769 (to 4 d.p.)	

Although this classifier uses *a posteriori* information with regard to the $p : n$ ratio, I feel it is fairer and more realistic to use this information than to use a wholly random binary classifier (ratio $1 : 1$). The true ratio that exists in Flickr would be ultimately discoverable if every view of a photo was recorded as well as any subsequent Favourite label information, and then made publicly available. However, this was not the case at the time of writing and so an informed estimate is made. I leave the task of a large-scale manual evaluation of users in which their Favourite labelling activity, both applying the label and otherwise, is monitored to future work.

[6] See Section 4.2.5 that describes the data sets for an explanation of this value.

4.2.4 Use-case scenarios

For machine learning tasks where completely labelled data is available, supervised techniques are suitable for training. If however, only some of the data is labelled, then semi-supervised techniques can be used. In the case of learning to identify Flickr photos labelled as Favourites, the Favourite photos themselves become the positive examples, however there is no easy analogy of a non-Favourite training example.

It is possible to train models using only positive and unlabelled examples (Elkan and Noto, 2008). Most techniques that accomplish this do so by using a heuristic to guess likely negative examples (or assign them weights) and then apply a standard learning algorithm.

This is in essence what I do here in that I use explicitly defined positive and pseudo-negative examples. While positive examples are easily identifiable (those labelled Favourite), for the pseudo-negative[7] examples I have come up with two alternative scenarios.

In the first scenario, I assume that the prior probability of a user having seen a certain photo follows a uniform distribution. This is not at all realistic, but it is simple to use and understand. In this scenario, the negative judgements for a user are sampled at random over all photos in Flickr that were not called a Favourite by any user. I refer to this scenario as the *Random Scenario*.

In reality, the social dynamics and interface design of Flickr, and other social media sharing sites, influence which photos a user is likely to be exposed to. This is reflected in the *Social Random Scenario*, where the set of *negative* judgements for each user is pooled at random from the non-Favourite photos belonging to that user's contacts and groups. This is more realistic, as there is a higher likelihood that the user has seen the photos and decided not to label them as Favourites.

[7] From this point on, when reference is made to *negative* examples, it should be assumed that this refers to the *pseudo-negative* examples defined here.

4.2.5 Datasets

In order to design my dataset, the usage of the Favourite label in Flickr must first be explored. Figure 4.2 shows the sampled distribution of the number of Favourite photos per user for all users in Flickr (as of May 2008) who use the label, using a log-log scale.

The x-axis represents 10,000 unique users, equally sampled from an ordered list of *all* Flickr users that have collected Favourite photos, sorted by descending number of Favourites. The y-axis refers to the number of Favourites each user has labelled. The distribution can be *approximated* by a power law (Reed, 2001), and the probability of a user having a Favourites frequency x is proportional to $x^{-1.1513}$. This is similar to the findings of Sigurbjörnsson and van Zwol (2008) where they showed that the distribution of textual tags used in Flickr (of which I suggest the Favourite label is one instance) also follows a power law.

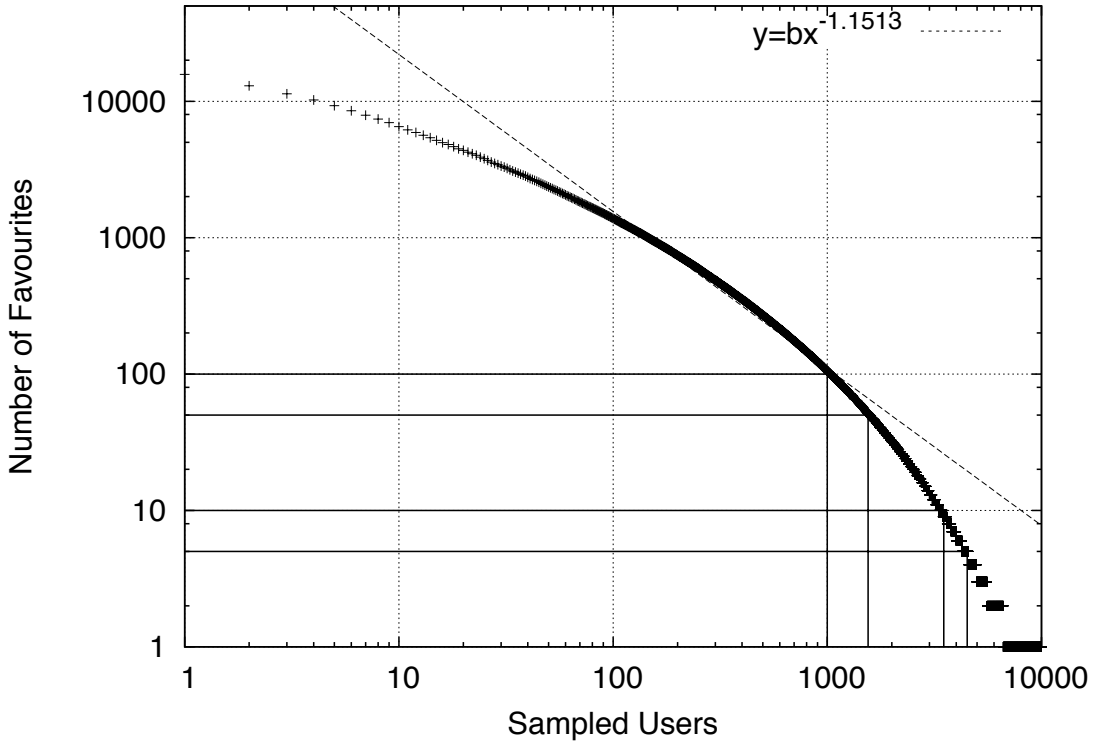


FIGURE 4.2: The sampled distribution of the number of Favourite photos per Flickr user

As can be seen from Figure 4.2 roughly 50% of the Flickr community have more than 5 photos marked as Favourites, and around 10% have more than 100. These values show a

distribution that I split into 4 sets to make it easier to analyse the effect of labelling activity on system performance, maintaining good coverage of the Flickr community: 5-9, 10-49, 50-99 and >100.

The positive examples used in this experiment are the Favourite images of users taken directly from the dataset. However, as mentioned in Section 4.2.4, there is no analogous label for non-Favourite images, and so I select negative examples in two ways, one of which is conceptually simple and fair, and another which is more realistic.

Random Scenario The dataset for the random scenario consists of the positive examples for that user set, complemented with photos randomly selected from throughout Flickr that have not been labelled as a Favourite by any Flickr user. The number of negative samples outnumbers the positive samples which reflects the experience of a typical user, who will only mark a small proportion of the photos they see as a Favourite. As a ratio between positive and negative examples had to be chosen for this experiment, and as there is no authoritative value to be found in the literature for this field, I had to chose a plausible value. I needed a value that reflected the distinct difference in usage but that still allowed me to train an effective classifier. From inspection of the data and initial testing, I chose a ratio between positive and negative samples per user of 1:6.5.

As no previous work has been able to determine an accurate value for this ratio in Flickr interaction, I pick this plausible value with future work in mind that will derive a more accurate value from user evaluation studies. For example, using a human interaction task market place like Amazon's Mechanical Turk[8], many users could be asked to evaluate a randomly-selected sample of images from Flickr and label them with a Favourite label should they wish to. Those images that aren't given that label can be counted and the ratio between Favourite and otherwise can be calculated.

[8] http://www.mturk.com/

Social Random Scenario While choosing images randomly from throughout Flickr is simple and unbiased, it does not accurately reflect the kind of interactions users have with the system. I speculate that many users browse through images from their contacts and that photo-sharing websites encourage this through notifications and reminders of contacts' activity. Therefore, for the *Social Random~* scenario, I gathered the same number of negative examples as the previous scenario, but this time the photos were pooled at random from the images of a user's social contact network, maintaining the same 1:6.5 ratio of positives-to-negatives. I ensured that none of the selected photos had been labelled as a Favourite by that user. However, this does mean that these photos may have been labelled as Favourites by other users.

Finally, both data scenarios have been randomly partitioned, using 70% of the data per user for training the classifier and the remaining 30% per user for testing the performance, while maintaining the 1:6.5 ratio of positive and negative labels for each user, as shown in Figure 4.3. As no parameter optimisation or model selection is carried out, no validation sets are required.

To allow comparison of the performance across the two datasets, I made sure that the test set for both scenarios contain the same positive examples.

The negative examples in the *Social Random~* scenario are likely to be more similar to the positive examples than in the *Random~* scenario. I base this on the assumption that those who are in close online social proximity to ourselves are likely to be more similar to us when compared to a completely random stranger from anywhere in the world. This would, however, make it more difficult to train a model that accurately classifies the test examples by discriminating between positive and negative examples. In this case, the performance of models trained for the *Social Random~* scenario would yield lower performance according to the metrics I selected for this experiment.

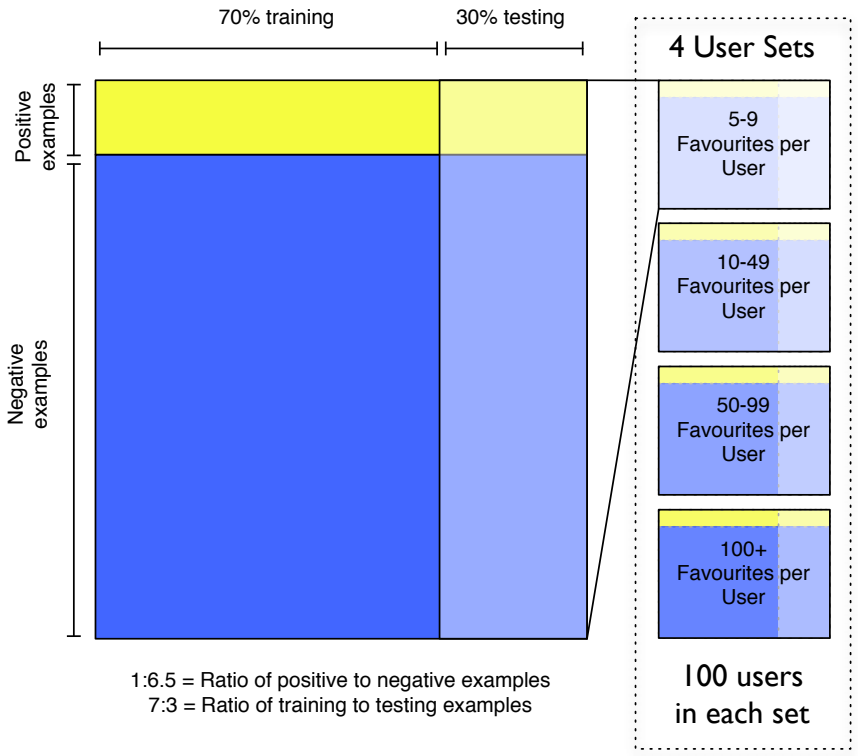


FIGURE 4.3: The partitioning of data between training and testing, for all user sets

4.2.6 Social features

The first class of features extracted from the images describes the connection between the photo owner and a user that has marked that photo as a Favourite. It also includes features that represent the connection between the groups in which a photo is posted, and the groups to which the user is subscribed. Table 4.1 describes the 10 social features derived for the experiment. They cover a range of different interactions types between the users and their photos and between the users themselves.

They include explicit connections between users as well as implicit connection information derived from the social graphs formed by users' interactions. In these social graphs, users are represented as nodes and the users' interactions are represented as edges between the nodes. Each relationship type can be used to form a different social graph (or, when multiple edges can exist between nodes, *multigraph*). For example, the *sharedGroups* feature is extracted from the social graph that connects users via shared membership of Flickr groups, similar

Table 4.1: Listing of all 10 social features

Feature	Description
viewsCount	The number of unique occasions the photo has been viewed.
contactsShared	The number of contacts that both the user and the owner of the photo share.
isContact	Binary indicator of whether the owner of the photo is a contact of the user or not.
groupsShared	The number of groups of which the user and owner of the photo share membership.
isFriend	Binary indicator of whether the owner of the photo is labelled a 'Friend' by the user.
isFamily	Binary indicator of whether the owner of the photo is labelled a family member by the user.
photoInGroup	The number of occasions the photo appears in a group of which the user is a member.
commentCount	The number of comments added to the photo.
uniqueCommenters	The number of unique users to have commented on the photo.
uniqueCommentsRatio	The ratio of unique commenters to total commenters on the photo.

to the *Social Group Context* defined in Chapter 3.

Membership of Flickr groups has a wide range of motivations, including sharing similar photos related to a common theme, sharing photos between members of certain demographics or sharing photos related to types of user behaviour. If two users, represented by nodes in our social graph, are members of the same group, an edge between their nodes describes this relationship. The *sharedGroups* feature counts the edges between the nodes of the viewer and the owner of a photo.

In a similar fashion, I extract information from the graph formed by the 'Contact' relationship in Flickr, that allows people to explicitly link themselves to other users, as a contact and additionally as a friend or family member. This multigraph is similar to that used in the tags suggestion experiment, described in Section 3.3.3. The *isContact* feature simply shows whether two users are each other's 'Contact', and the *isFriend* and *isFamily* features describe 'Friend' and 'Family' links, respectively. For the *contactsShared* feature, I use the order of the set described by the intersection of the neighbourhoods in the 'Contact' graphs of the user in question viewing a photo and the owner of that photo, giving the number of contacts

Table 4.2: Listing of all 25 textual features

Feature	Description
tagCount	Total number of tags of given photo
photosTagsProb_{min, max, mean, variance}	Statistics of the photo tags w.r.t the Flickr tag space
favouritesTF_{min, max, mean, stdDev, cosSim}	Statistics of tag TF values w.r.t. user's Favourite photos
favouritesTFIDF_{min, max, mean, stdDev, cosSim}	Statistics of tag TFIDF values w.r.t. user's Favourite photos
userUploadsTF_{min, max, mean, stdDev, cosSim}	Statistics of tag TF values w.r.t. user's uploaded photos
userUploadsTFIDF_{min, max, mean, stdDev, cosSim}	Statistics of tag TFIDF values w.r.t. user's uploaded photos

they have in common. I propose that the higher the overlap between the neighbourhoods of two users, the more likely they are to be closely connected.

I also look at more indirect relationships between users and photos by measuring how many groups that a user is a member of contain the image in question with the *photoInGroup* feature. A higher number suggests that the image matches well with the themes of the user's groups. As a general measure of interest in a photo, I collect the number of views an image receives using the *viewsCount* feature. The more often an image is viewed, the more popular the image is likely to be among Flickr users in general.

The issue of temporal granularity for such features was not addressed in this experiment, but will be looked into in future work, as will temporal features in general. These temporal features could include detecting clusters and repeatable patterns of interactions and events, as well as measures that describe 'freshness' of data.

4.2.7 Textual features

I use a total of 25 textual features, as shown in Table 4.2. The four *photosTagsProb* features are derived from the tags of the photo and are based on the probability of a given tag occurring in a photo uploaded to Flickr, computed with respect to the entire Flickr tag space as of May 2008. I then calculate the minimum, mean, maximum and variance of these probability values for the group of tags associated with the photo in question. I also add the simple total number of tags for that photo as a feature. This gives 5 features per photo.

Next I calculate features that are dependent on knowing the user for whom judgements are being made. For that I use the classical vector space model for text retrieval (Salton et al., 1975). This involves comparing the tags of the photo in question against a set of tags that represent the user. I use two such aggregations:

- The first is the aggregation of all the tags associated with all the Favourite images of the observing user. This captures a sense of the topical nature of the user's interests.
- The second is the aggregation of all the tags associated with all the images uploaded by the user. This represents the vocabulary of the user.

I then compute the minimum, mean, maximum and standard deviation of the term frequency (TF) and the term frequency $\times$ inverse document frequency (TF-IDF) values of the tags of the photo with respect to both aggregations.

For each aggregation I also compute the cosine similarity between the photo's tag TF and TF-IDF values. This reflects how well the photo's tags match the user's past tagging behaviour, giving an additional 4 features.

4.2.8 Visual features

The objective of the visual features is to capture the perceivable nature of an image. This may be at a simple level (size, dominant colour, etc.), the calculation of which tends to be computationally cheap to compute for the size of photos commonly uploaded online, or at a higher level that attempts to capture the aesthetics of an image, but usually at a higher computation cost.

I chose a set of low-dimensional image features that describe attributes that encapsulate the human perception of the photos, rather than more high dimensional features like colour histogram, edge directionality (CEDD), and other well known global image features. San Pedro and Siersdorfer (2009) have shown that such features perform well when classifying images

Table 4.3: Listing of all 39 visual features

Feature	Description
Orientation	Width/Height ratio
Size	Pixel count
Contrast	Score
RMSContrast	Score
Saturation	{Min, Max, Avg, StdDev} contrast values
Brightness	{Min, Max, Avg, StdDev} contrast values
Sharpness	Score
Colourfulness	Score
Sky	{Proportion, Score} of sky colours
Vegetation	{Proportion, Score} of vegetation colours
Skin	{Proportion, Score} of skin colours
Naturalness	Combined score of sky, veg. and skin
Tamura	18 dimension texture feature

based on general attractiveness, which I consider as a task similar to that of Favourite recommendation.

Some of the following features are colour-space agnostic, others depend on handling an image in a particular representation. All images in the dataset were stored in Cartesian RGB (Red, Green, Blue) representation, but were converted to the cylindrical-coordinate representations HSL (Hue, Saturation and Lightness/Luminance) or HSV (Hue, Saturation and Value) when required. The MPEG-7 Color and Texture Descriptors paper by Manjunath et al. (2001) provides a useful overview of colour spaces with respect to image features extraction (see Section 2.5).

The full set of visual features is outlined in Table 4.3. In total 39 visual features are used.

4.2.8.1 Geometry

The first of the two geometric features describes orientation and indicates whether an image is portrait or landscape, defined as the ratio of height versus width in pixels. The second is a size feature that counts total pixels and that allows for differentiation between high resolution images taken with cameras of potentially higher quality (DSLRs for example), and images taken with smaller sized cameras or other mobile devices that may lack the lens, sensor and digital signal processing capabilities of the former.

4.2.8.2 Contrast

I compute two types of contrast features, known here as Contrast and the normalised Root Mean Square (RMS) Contrast. For the first, the image is converted to an HSL representation and the average distance between the luminance of each pixel $l_{x,y}$ of total N pixels is calculated, as well as the average image luminance $\bar{I}$:

$$C = \frac{1}{N} \sum_{x,y} \left(l_{x,y} - \bar{I} \right) \tag{4.3}$$

RMS Contrast (C_{RMS}) allows for fairer comparison between independent images and is computed by first calculating an average normalised image luminance $\bar{L}$:

$$\bar{L} = \frac{1}{N} \sum_{x,y} \frac{l_{x,y} - l_{min}}{l_{max} - l_{min}} \tag{4.4}$$

$$C_{RMS} = \sqrt{\frac{1}{N} \sum_{x,y} \left(l_{x,y} - \bar{L} \right)^2} \tag{4.5}$$

4.2.8.3 Saturation, brightness, sharpness and colourfulness

The saturation, brightness and colourfulness features describe the colour characteristics of the image in terms of minimum, average, maximum and standard deviation of vividness and luminance, and a score for difference-from-grey respectively, giving 9 colour based features. Saturation is most easily obtained in a colour space that uses it as one of coordinates to describe a shade—HSV for example. To save colour-space conversion, I calculate:

$$\text{Sa} = \max(R, G, B) - \min(R, G, B), \tag{4.6}$$

where R, G and B are the colour values of a pixel in the sRGB[9] colour space, for all pixels, then the statistics are calculated over all the pixel saturation values.

I define brightness as the average intensity of all the pixels in the image. Again, using a colour space that encodes luminance directly, its calculation in the YUV colour-space is the mean over all pixels:

$$\bar{Y} = \frac{1}{N} \sum_{x,y} (Y_{xy}), \tag{4.7}$$

where Y_{xy} describes the luminance value for a pixel at coordinates x, y and N is the total number of pixels. Statistics are then computed for all pixel brightness values.

Sharpness measures the coarseness of the image and can be determined as a function of its discrete Laplacian, using normalised local average luminance with respect to the surrounding pixels.

$$\text{Sh} = \sum_{x,y} \frac{L(x,y)}{\mu_{x,y}} \tag{4.8}$$

$$L(x,y) = \lim_{\epsilon \to 0} \frac{[F(x+\epsilon) - F(x)] + [F(x-\epsilon) - F(x)]}{\epsilon^2} \tag{4.9}$$

where $\mu_{x,y}$ is the average luminance of the pixels around pixel of coordinates x, y of image I. Statistics are then computed for al pixel sharpness values.

Colourfulness (Cf), as defined by Hasler and Suesstrunk (2003) can be extracted in the sRGB colour space using a derivative opponent colour space defined as:

[9] sRGB is a parameterised instance of the RGB colour space that uses primary colour points as defined by the International Telecommunication Union in their recommendation standard ITU-R BT.709

$$rg = R - G \tag{4.10}$$

$$yb = \frac{1}{2}(R + G) - B \tag{4.11}$$

Colourfulness is then calculated as:

$$\text{Cf} = \sigma_{rgyb} + 0.3 \cdot \mu_{rgyb} \tag{4.12}$$

$$\sigma_{rgyb} = \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2} \tag{4.13}$$

$$\mu_{rgyb} = \sqrt{\mu_{rg}^2 + \mu_{yb}^2} \tag{4.14}$$

4.2.8.4 Naturalness

The naturalness feature I use is first suggested by Huang et al. (2006) and attempts to combine multiple aspects of the visual perception of an image including colourfulness and dynamic range into a single score. This score is made up of three constituent parts representing the proportion of pixels judged likely to be either human skin, grass-like vegetation or sky. Using the HSL colour-space, the pixels are first filtered for those that have values $20 \leq L \leq 80$ and $S > 0.1$. The remaining pixels are then grouped in 3 sets: Skin, Grass and Sky according to the hue ranges shown below. The average saturation μ_S for each group is used to compute naturalness indexes for each group:

$$N_{\text{Skin}} = e^{-0.5\left(\frac{\mu_S^{\text{Skin}} - 0.76}{0.52}\right)^2}, \text{if } 25 \leq \text{hue} \leq 70$$
$$N_{\text{Grass}} = e^{-0.5\left(\frac{\mu_S^{\text{Grass}} - 0.81}{0.53}\right)^2}, \text{if } 95 \leq \text{hue} \leq 135$$
$$N_{\text{Sky}} = e^{-0.5\left(\frac{\mu_S^{\text{Sky}} - 0.43}{0.22}\right)^2}, \text{if } 185 \leq \text{hue} \leq 260$$

These are then combined to form a score for the image's naturalness:

$$N = \sum_i \omega_i N_i, \; i \in \{\text{'Skin'}, \text{'Grass'}, \text{'Sky'}\} \tag{4.15}$$

where ω_i is the proportion of pixels in group i with respect to the total pixels in the image.

I use the scores for each of the three pixel group as well as the overall naturalness score as features. I also include the proportions of each pixel type with respect to the total image pixels. This gives 7 naturalness features in total.

4.2.8.5 Texture

Tamura features characterise the texture of the image using coarseness, contrast and directionality, as described in the work of Tamura et al. (1978). Coarseness and contrast are represented as single numeric values whereas directionality is a 16 bin histogram. This therefore gives 18 values that represent the Tamura texture of the image.

4.2.9 Implementation

Based on the data partitioning method described in Section 4.2.5, Table 4.4 shows the scale of my dataset in general and for each of the four specific sets. Each set has 100 unique users. The two negative example scenarios have slightly different numbers of examples due

Table 4.4: Total number of examples (both for training and testing) for each group of 100 users within each set

Range	Number of users	Positive instances	Negative instances	
			Social Random	Random Scenario
5-9	100	671	25,961	25,881
10-49	100	2,211	14,487	14,420
50-99	100	6,947	45,275	45,210
>100	100	63,325	411,637	411,682

to the probabilistic gathering algorithm used that maintained the correct ratios of positive-to-negative for each user, at the expense of producing identically sized sets.

The users with 100 or more Favourites provide a mean average of 633.25 positive examples per user with respect to my sample, but ultimately only represent around 10% of the greater Flickr community.

The users with smaller ranges of Favourites (5-9, 10-49, 50-99) represent approximately 90% of Favourites-using Flickr users—as can be seen in Figure 4.2—and are a significant proportion of the total population, even though they are less active in using the label.

All four sets are divided between test and training data as shown in Figure 4.3 in Section 4.2.5. The textual, visual and social features were then extracted.

Training gradient boosted decision tree. I trained the GBDTs as described in Section 4.2.2 with a maximum of 300 sub-trees, 30 leaf nodes per tree, with a learning rate of 0.04 and a sampling rate of 0.5. These values were derived through analysis of existing related work and initial experimentation.

Experimental runs For the experiment, there are 14 runs, 7 for each data scenario: the *Random* and *Social Random*. These seven runs are based on the complete set of possible combinations of the features classes: *textual*, *visual*, and *social*. These 14 runs are repeated for each of the 4 user datasets, giving a total of 56 runs.

4.3 General classifier evaluation

The results from this part of my experimental analysis are summarised in Section 4.3.2.

This section presents a breakdown of results as a series of graphs dealing with specific subsets of users, which features were found to be particularly powerful in making predictions and how the importance of the full feature catalogue varied between these sets of users.

4.3.1 Overall performance

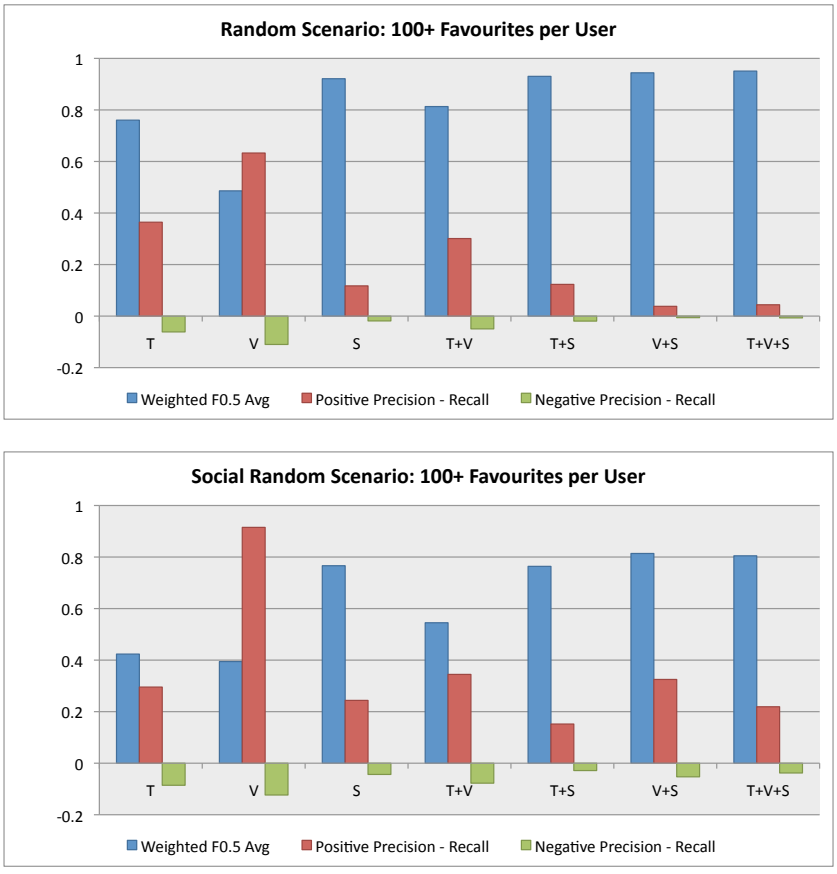


FIGURE 4.4: Performance for users with 100+ favourite images for both data scenarios, for textual (T), visual (V) and social (S) features. Full data found in Table 5.1 in the Appendices

The following analysis evaluates performance of runs by first looking at those of users who have '100+ Favourites'. In this case, the machine learnt classifier has more data per user with which to train (and test against) and so I suggest it is these runs that will give the clearest

(a) Random Scenario

Rank	Features	Relative Importance
1	SOCIAL_CONTACT	100
2	SOCIAL_CVIEWS	66.5501
3	VISUAL_NATURALNESS	40.4898
4	SOCIAL_CONTACTS_SHARED	20.9819
5	SOCIAL_PHOTO_IN_GROUP	19.1253
6	SOCIAL_UNIQUE_COMMENTS	9.08966
7	VISUAL_VEGETATION_SCORE	8.1726
8	VISUAL_SATURATION_AVG	7.83377
9	SOCIAL_COMMENTS_COUNT	7.30045
10	TEXTUAL_FAVOURITESTFIDF_MIN	6.57924

(b) Social Random Scenario

Rank	Features	Relative Importance
1	SOCIAL_CONTACT	100
2	SOCIAL_CVIEWS	56.7734
3	SOCIAL_GROUPS_SHARED	48.965
4	SOCIAL_CONTACTS_SHARED	30.7282
5	SOCIAL_UNIQUE_COMMENTS	30.6345
6	SOCIAL_PHOTO_IN_GROUP	28.6858
7	TEXTUAL_FAVOURITESTF_MAX	27.7613
8	TEXTUAL_FAVOURITESTFIDF_MAX	27.3057
9	SOCIAL_COMMENTS_COUNT	23.8886
10	TEXTUAL_USERUPLOADSTFIDF_CS	22.7892

Table 4.5: Top 10 most important features for runs with users who had 100+ Favourites, the distribution is graphed in Figure 4.5

picture of the systems' performance for active users. The three remaining runs (50-49, 10-49 and 5-9 Favourites per user) are then discussed in descending order of the number of Favourites.

This sequence will show how performance changes as the system addresses users with less Favourite labelling activity, who represent a bigger proportion of the Flickr community.

The following explanation of the graphs representing system performance holds for all runs.

4.3.1.1 Results for user with 100+ Favourites

Representing around 10% of Flickr users who have Favourites

Figure 4.4(b) shows the results for all fourteen runs for users with 100 or more Favourite images. The two different scenarios are shown in the graph. The values for the weighted

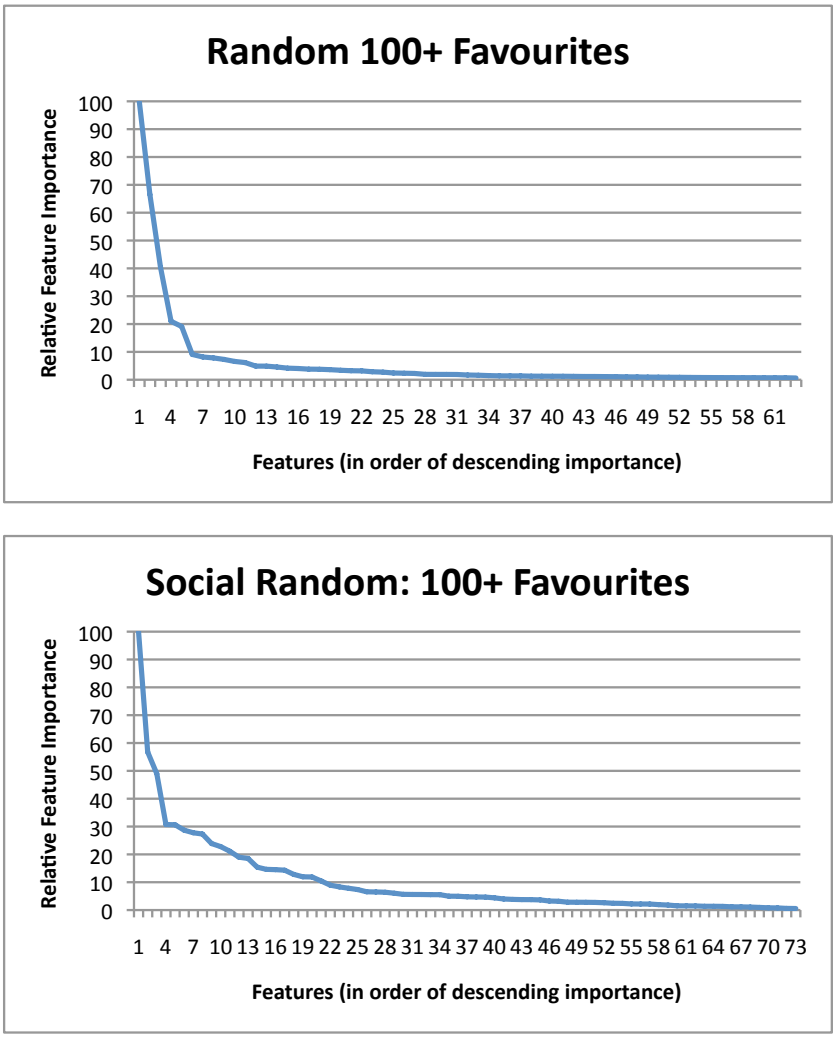


FIGURE 4.5: Distribution of feature importance over all features selected by GBDT for users with 100+ Favourites. The top 10 are shown in Table 4.5

$F0.5_{avg}$ measure are given as a general performance indicator, combining both precision and recall values[10] for both classes, weighted by the ratio of positive to negative examples (1:6.5). However, while the $F0.5_{avg}$ measure gives an impression of overall performance, it hides the relationship between precision and recall for the two classes, and this is valuable information when comprehensively evaluating these results. The bars marked "Positive Precision - Recall" show the difference between the precision and recall values for the positive class—examples labelled as Favourites, where larger bars indicate a larger difference. Similarly, the "Negative Precision - Recall" bars show the same for the negative class (examples that are not Favourites). These difference values give an indication as to whether it is

[10] Precision and recall are not shown here for the sake of simplicity, but all values are shown in the Tables in the Appendices, in this case, Table 5.1.

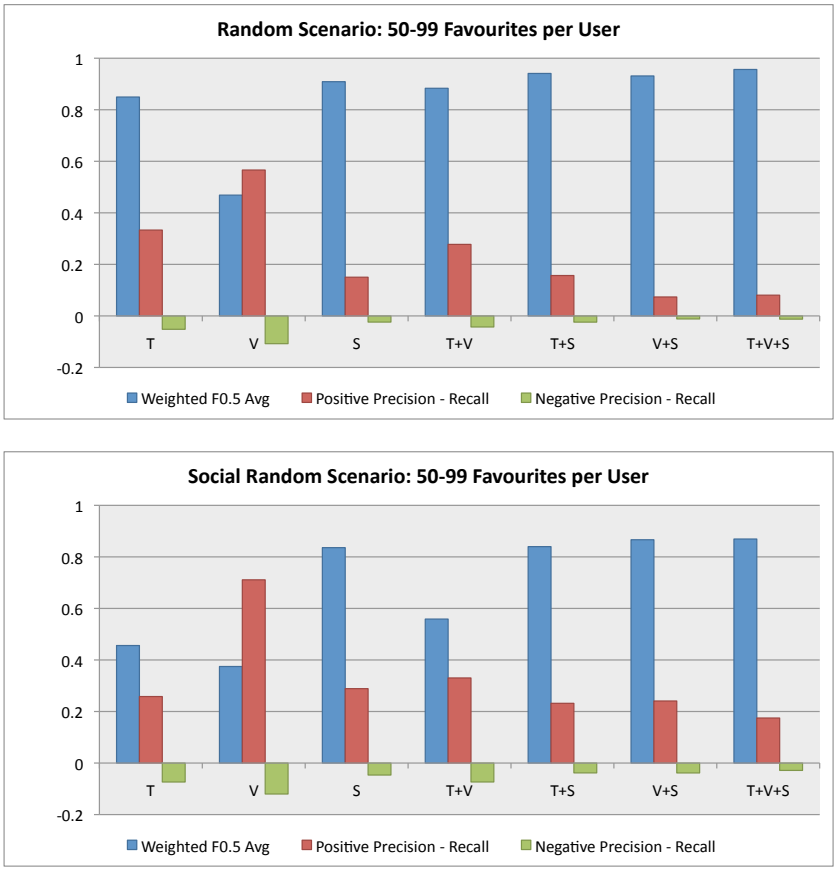


FIGURE 4.6: Performance for users with 50-99 favourite images for both data scenarios, for textual (T), visual (V) and social (S) features. Full data found in Table 5.2 in the Appendices

precision or recall that dominates the F-Measure with respect to each class.

In Section 4.2.3 I presented a trivial probabilistic classifier, the performance of which can be used to give context to the experimental classifier results. By comparing its expected precision, recall and $F0.5_{avg}$ to the experimental results (see Appendix **??**), it can be seen that in terms of precision of the Favourite class, all models were above performance of the trivial classifier. Recall for the Favourite class however was below or very close to the trivial classifier for visual only runs for all users and in both scenarios.

For non-Favourites, all runs exhibited higher performance than the trivial classifier for both precision and recall. In terms of the summary metric $F0.5_{avg}$, in the *Random* scenario, the textual and visual runs were close to, or below the trivial classifier. In the *Social Random* scenario, the textual, visual and the combination of textual and visual runs tended to be below the trivial classifier.

(a) Random Scenario

Rank	Features	Relative Importance
1	VISUAL_VEGETATION_SCORE	100
2	VISUAL_SKIN_PROPORTION	57.7853
3	VISUAL_NATURALNESS	55.8274
4	SOCIAL_CVIEWS	55.0558
5	VISUAL_SKIN_SCORE	51.764
6	VISUAL_SATURATION_STDDEV	48.5806
7	SOCIAL_CONTACT	23.9515
8	VISUAL_SKY_SCORE	16.2099
9	TEXTUAL_FAVOURITESTFIDF_CS	13.7097
10	VISUAL_BRIGHTNESS_AVG	11.0467

(b) Social Random Scenario

Rank	Features	Relative Importance
1	SOCIAL_CONTACT	100
2	SOCIAL_CVIEWS	82.3686
3	TEXTUAL_FAVOURITESTFIDF_MAX	42.2399
4	VISUAL_BRIGHTNESS_AVG	42.1779
5	SOCIAL_UNIQUE_COMMENTS	34.1285
6	SOCIAL_GROUPS_SHARED	31.8776
7	SOCIAL_COMMENTS_COUNT	27.797
8	SOCIAL_PHOTO_IN_GROUP	20.7267
9	VISUAL_SKY_SCORE	19.1851
10	SOCIAL_CONTACTS_SHARED	18.288

TABLE 4.6: Top 10 most important features for runs with users who had 50-99 Favourites

Overall, this reinforces the finding that visual by itself is a bad source for recommendations when used entirely by itself,and that the textual data is also not, by itself, very useful for this particular task. Again it was found that it was harder to discriminate between the classes in the Social Random scenario.

Across feature combinations By first looking at the *Random Scenario* it can be seen that the three individual feature class runs (Textual (T), Visual (V) and Social (S)) show three distinctly different performance levels with respect to Weighted $F0.5_{avg}$. The Visual run is particularly poor. However, by examining the Precision - Recall difference values, it can be seen that it is the bad performance in terms of recall that dominates. So while visual features may provide results that are precise (0.781), this class of features has trouble extracting possible examples of Favourite images from the test collection. The Textual run performs

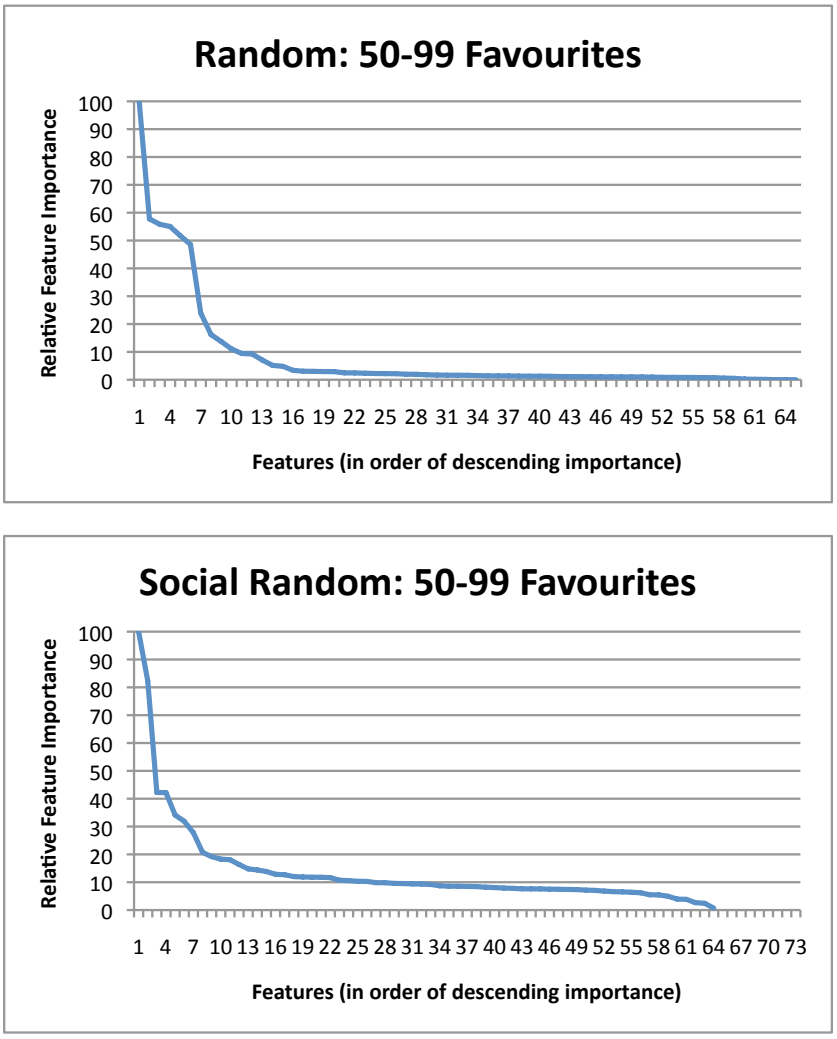


FIGURE 4.7: Distribution of feature importance over all features selected by GBDT for users with 50–99 Favourites.

well, with a smaller difference in precision and recall. But it is the Social run that clearly outperforms the others, in terms of both precision and recall.

The "Negative Precision - Recall" values are small for all runs, but are more pronounced for the Visual and Textual runs. This shows that accurately classifying non-Favourite images is easier than accurately classifying Favourites.

Across scenarios The *Social Random* scenario is very similar to the *Random* scenario in terms of relative performance between classes and combinations, but it performs consistently worse when values are compared between the scenarios themselves. However, the

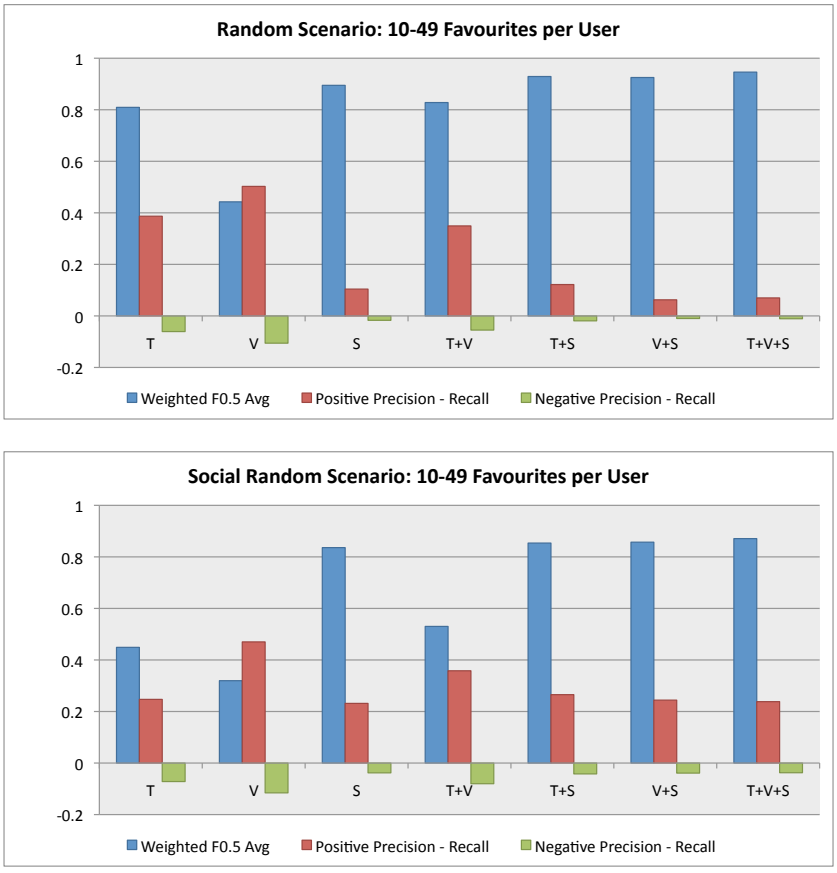


FIGURE 4.8: Performance for users with 10-49 favourite images for both data scenarios, for textual (T), visual (V) and social (S) features. Full data found in Table 5.3 in the Appendices

following differences can be observed:

- Except for the Textual run, most runs saw a reduction of around 10% in Weighted $F0.5_{avg}$. The Textual run saw a decrease of 44.3%. This seems to suggest that the features used in the Textual runs were more sensitive to the design of the negative example sets than the other features classes. This pronounced reduction in performance is also found in the combination runs where Textual features were used, although to a lesser degree.

- The Visual run has a particularly high difference between precision and recall, and after inspection of the data, it can be seen that while this run had 100% precision, it had a very low recall of 0.085. This is also reflected in the relatively low recall for negative examples of 0.394.

(a) Random Scenario

Rank	Features	Relative Importance
1	VISUAL_SATURATION_STDDEV	100
2	VISUAL_SATURATION_AVG	44.4364
3	VISUAL_VEGETATION_SCORE	31.7865
4	SOCIAL_CVIEWS	29.2692
5	VISUAL_SKIN_SCORE	19.4846
6	SOCIAL_CONTACT	12.8611
7	TEXTUAL_FAVOURITESTF_MEAN	4.94248
8	TEXTUAL_FAVOURITESTFIDF_MAX	4.8133
9	SOCIAL_COMMENTS_COUNT	4.49066
10	TEXTUAL_FAVOURITESTFIDF_CS	3.83087

(b) Social Random Scenario

Rank	Features	Relative Importance
1	SOCIAL_CONTACT	100
2	SOCIAL_CVIEWS	90.0416
3	TEXTUAL_FAVOURITESTFIDF_MAX	45.1829
4	SOCIAL_GROUPS_SHARED	38.6336
5	VISUAL_BRIGHTNESS_AVG	37.0564
6	SOCIAL_UNIQUE_COMMENTS	35.9525
7	SOCIAL_COMMENTS_COUNT	34.9313
8	TEXTUAL_FAVOURITESTFIDF_CS	25.4096
9	TEXTUAL_FAVOURITESTFIDF_MEAN	24.6093
10	TEXTUAL_FAVOURITESTF_MEAN	24.0943

Table 4.7: Top 10 most important features for runs with users who had 10-49 Favourites

- The difference between positive class precision and recall is more pronounced in general for all runs.
- The negative class difference in precision and recall is very similar, even though the positive class varies.

Ultimately the results suggest that changing the negative example selection method has a significant impact on the classifiers' performance on positive results, while the negative results are generally similar. This could be partly explained by the ratio between positive-to-negative examples used for training. Any change to the nature of the majority of examples (in this case, non-Favourites) could disproportionately effect the performance on the minority.

It also provides confirmation of my initial hypothesis regarding the similarity of examples in the positive and negative classes as defined in Section 4.2.5. The generally lower performance

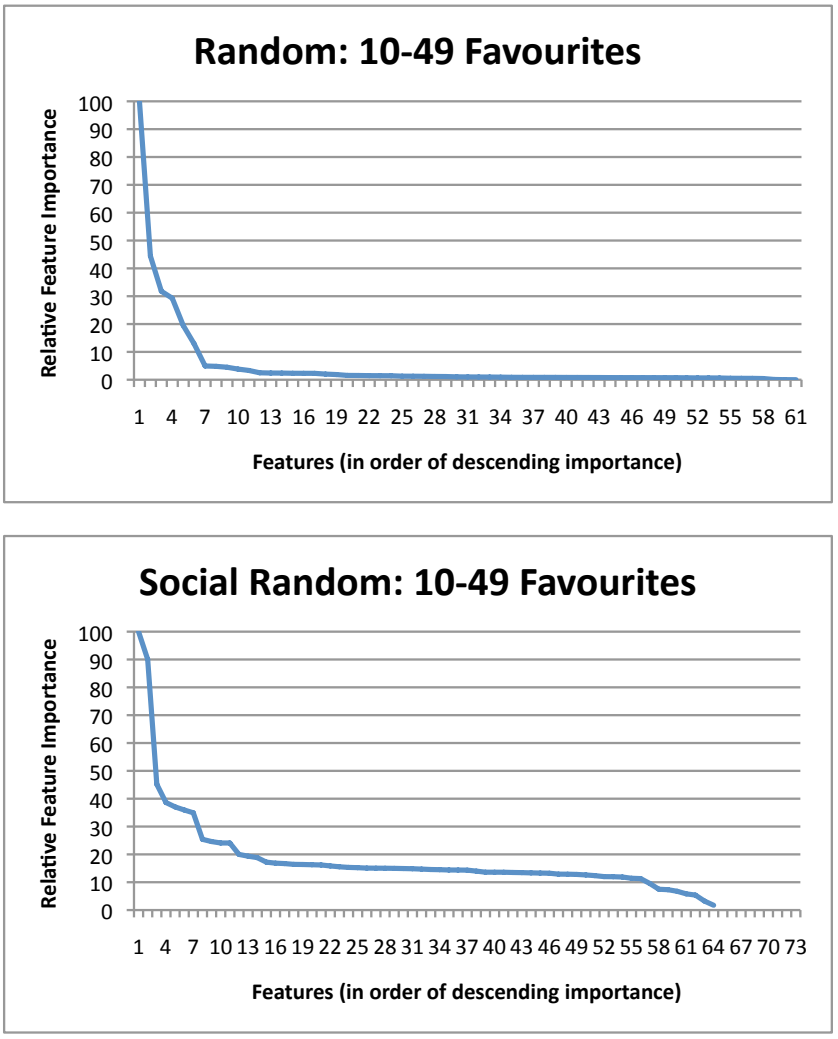


FIGURE 4.9: Distribution of feature importance over all features selected by GBDT for users with 10-49 Favourites.

of the models in the *Social Random~* scenario would indicate that the examples are harder to distinguish and hence are more similar.

Best overall performance The best performing run with respect to Weighted $F0.5_{avg}$ varied between the two scenarios. For the *Random~* scenario, the combination of all three feature classes (T+V+S) outperformed all others. However, in the *Social Random~* scenario, the Visual + Social run was best. I suggest that this could be due to the poor Textual results for this scenario mentioned above actively degrading the overall performance. Without this harmful addition, the V+S run is slightly better.

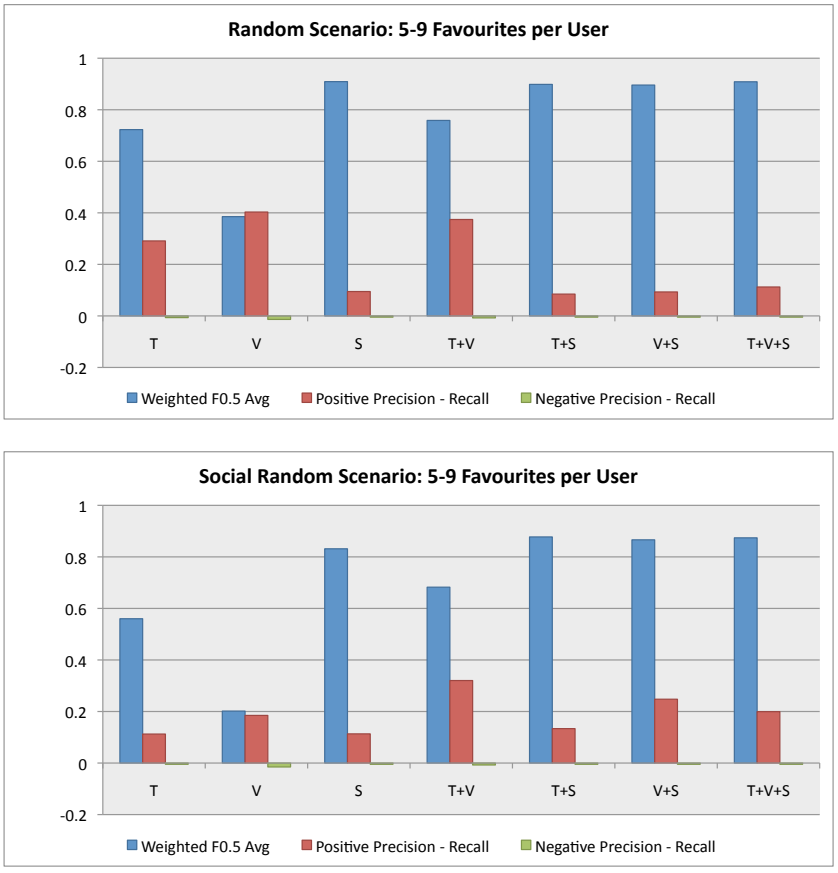


FIGURE 4.10: Performance for users with 5-9 favourite images for both data scenarios, for textual (T), visual (V) and social (S) features. Full data found in Table 5.4 in the Appendices

Across classes It is immediately evident that the metrics for the non-Favourite class are similar to or distinctly higher than those for the Favourite class, across the different combinations of feature classes. This implies that it is far easier to identify images that are not going to interest the users, than those that are more likely to. Most interestingly, this is even the case when using only visual features, meaning that within the use-case of this experiment, images could still be usefully discarded from consideration as Favourites without requiring textual metadata or social context information. (although to a lesser degree, e.g. there is a 7.35% reduction in $F0.5$ between the non-Favourites in the Social run and the Visual run). However this is to a lesser degree, e.g. the Visual run has the lowest $F0.5$ for non-Favourites in both scenarios, with a difference between the best run for this metric (T+V+S) of 8.4% in the *Random~* scenario and 0.52% in the *Social Random~* scenario.

(a) Random Scenario

Rank	Features	Relative Importance
1	SOCIAL_CVIEWS	100
2	SOCIAL_CONTACT	31.5672
3	SOCIAL_COMMENTS_COUNT	15.9298
4	TEXTUAL_USERUPLOADSTFIDF_CS	12.9258
5	TEXTUAL_FAVOURITESTFIDF_MAX	10.9894
6	VISUAL_SKY_PROPORTION	10.7133
7	TEXTUAL_FAVOURITESTF_MEAN	10.1611
8	TEXTUAL_FAVOURITESTFIDF_CS	9.35092
9	VISUAL_SKIN_SCORE	9.23745
10	TAGS_MAX	8.85552

(b) Social Random Scenario

Rank	Features	Relative Importance
1	SOCIAL_CVIEWS	100
2	SOCIAL_CONTACT	82.3449
3	TEXTUAL_USERUPLOADSTFIDF_CS	34.5847
4	SOCIAL_CONTACTS_SHARED	31.8625
5	TEXTUAL_USERUPLOADSTFIDF_MIN	18.2466
6	TEXTUAL_FAVOURITESTFIDF_MAX	16.8504
7	SOCIAL_COMMENTS_COUNT	15.6702
8	VISUAL_SKY_PROPORTION	15.4452
9	SOCIAL_GROUPS_SHARED	15.1261
10	VISUAL_TAMURA_19	14.7977

Table 4.8: Top 10 most important features for runs with users who had 5-9 Favourites

Being able to correctly identify non-Favourites is a valuable ability considering how relatively little users annotate their images.

Feature importance Table 4.5 shows the top ten features used by the GBDT in descending order of their relevance to the classifier as calculated by equations (39-41) in Friedman (2001) on greedy function approximation. The most influential feature (the one that gives the greatest empirical improvement in squared-error as a result of tree branching) is arbitrarily given a value of 100. All others have scaled values relative to this most important feature.

This ultimately produces a relative ranking for all features used in the model. As the importance values are relative to the most important feature in that run, values cannot be directly compared between runs, only their relative rank positions.

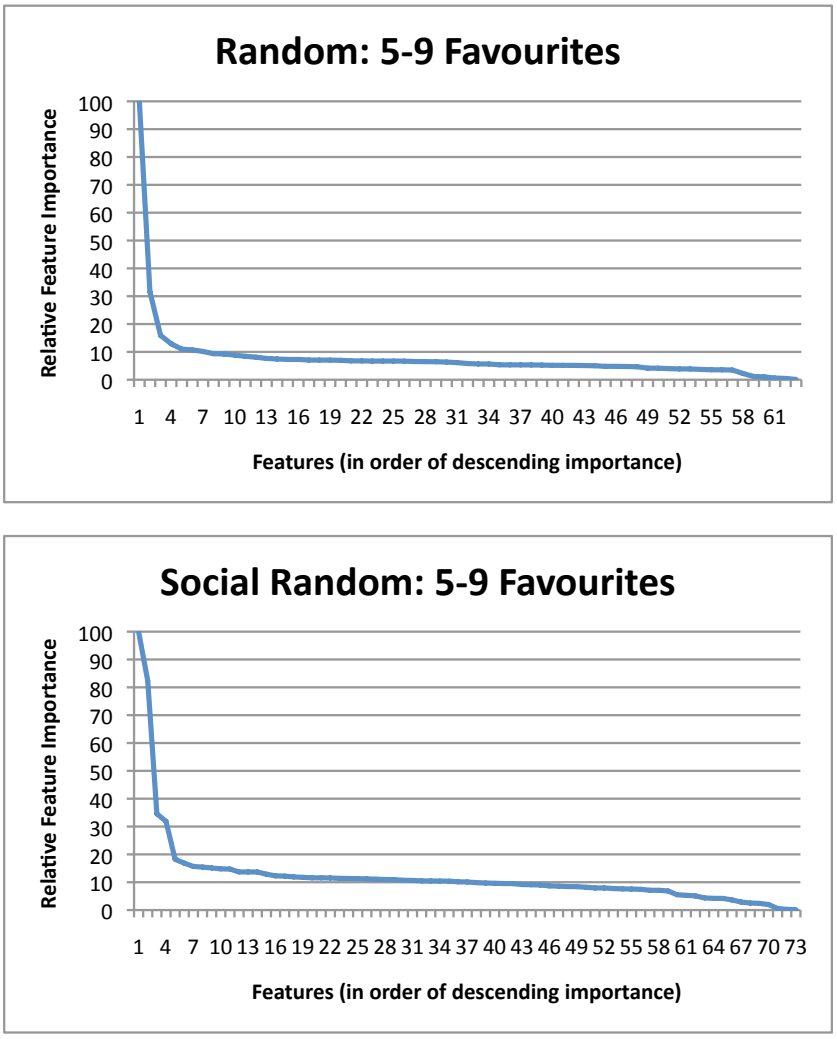


FIGURE 4.11: Distribution of feature importance over all features selected by GBDT for users with 5–9 Favourites.

For users with 100+ Favourites it can be seen that for the *Random* scenario in Table 4.4(a), social features dominate the top 10 most important features and that the visual naturalness feature is also very useful. This supports my hypothesis that features that describe social connections between users, and between users and their media are important in image recommendation and that multiple classes of features can usefully combine to provide better results than individual classes.

The only textual feature in the top 10 is the minimum of the TF-IDF values generated by the favourites of the viewing user with respect to the candidate image.

As a summary of the top most relatively important features for each run, Table 4.9 shows

Table 4.9: Overview of feature importance for each bucket of users

Number of Favourites	Importance Rank	*Random~* Scenario	*Social Random~* Scenario
100+	1	Social Contact	Social Contact
	2	Social Views Count	Social Views Count
	3	Visual Naturalness	Social Groups Shared
50-99	1	Visual Vegetation Score	Social Contact
	2	Visual Skin Proportion	Social Views Count
	3	Visual Naturalness	Textual Favourites TFIDF Max
10-49	1	Visual Saturation StdDev	Social Contact
	2	Visual Saturation Avg	Social Views Count
	3	Visual Vegetation Score	Textual Favourites TFIDF Max
5-9	1	Social Views Count	Social Views Count
	2	Social Contact	Social Contact
	3	Social Comments Count	Textual User Uploads TFIDF Cosine Similarity

the top three features for each set of users and for each scenario. The three most highly occurring features throughout all runs are the Social Contact, Social Views Count and components of the Visual Naturalness feature. This is particularly interesting when considered against the general performance of the visual runs across all users and both scenarios, as the runs using just visual features tend to be very badly performing when compared to others (and the probabilistic trivial classifier).

The importance of features changes between the two scenarios, and the rank for the *Social Random~* scenario in Table 4.4(b) shows that while social features still dominate (and the same features for both scenarios), the formerly highly important visual features drop out of the top ten, with the visual naturalness feature dropping to rank position 39.

This seems to suggest that the difference in the negative examples has a significant effect on the most useful features for training the GBDT model. For examples taken completely randomly from throughout Flickr, visual features are useful, but when these images are taken from a user's contacts, the social and textual features play a more significant role.

Figure 4.5 plots the relative importance of all the features used by the GBDT for both scenarios. The x-axis shows the features in descending importance order. These graphs show how importance is distributed among the complete set of features used[11]. In this case, it can be seen that for the *Random~* scenario, most feature importance was concentrated in the first few features, whereas for the *Social Random~* scenario importance is spread deeper into the tail of the distribution. This variation between the two contexts is shared between all the four user sets used in the experiment. This would seem to imply that for the *Random~* scenario, the model is mostly dependant on the first few highly important features, but for the *Social Random~* scenario with a different negative example selection method, these features become relatively less useful and the remaining features become relatively more influential.

4.3.1.2 Results for user with 50-99 Favourites

Representing around 16% of Flickr users who have Favourites

As the labelling activity of the users is relaxed from the high 100+ range to 50-99 Favourites per user, we see only slight changes in performance over all metrics. However, Table 4.6 shows the top ten features for this run and it can be seen that between the 100+ and 50-99 user sets, the most important features vary considerably. For the 50-99 runs, visual features dominate in the *Random~* scenario and have greater importance in the *Social Random~* scenario. Social features are still important in both.

As for the 100+ user set, the distributions of feature importance for the 50-99 user set differ, with the *Random~* scenario concentrating most of the feature importance in the first feature (in this case the visual vegetation score) and around the next 5, whereas in the *Social Random~* scenario, more feature importance is found in the tail of the distribution.

[11] It should be noted that the feature distribution graphs in this chapter vary in the number of features they present due to the selection process inherent to the GBDT training method.

4.3.1.3 Results for user with 10-49 Favourites

Representing around 13% of Flickr users who have Favourites

Shifting the labelling activity window further downwards to those users with 10-49 Favourites, again only marginal changes can be seen, mostly minor degradation in performance from the 100+ to the 50-99 set. However, the combination of all three feature classes emerges as the consistently mostly highly performing for all metrics in both data scenarios (most easily seen in Table 5.3). The relative feature importance distribution is again similar to the previous two user sets in that most importance is found in the first few features and that the *Social Random* scenario has more importance in the tail of the distribution than the *Random* scenario.

4.3.1.4 Results for user with 5-9 Favourites

Representing around 10% of Flickr users who have Favourites

From Figure 4.10 the most striking result is how small the difference between precision and recall is for the negative examples. According to the data table in the Appendix, this is because these two metrics are both very high. The other metrics are similar to the other user set results, with only minor degradation in performance across the feature classes and combinations. For example, $F0.5$ is reduced by an average of only 4.2% between all *Random* scenario runs between the 10-49 dataset and the 5-9 dataset.

4.3.2 Summary of general classifier approach

The findings of the single classifier for all users approach can be summarised thus:

- While most features ultimately contribute to the trained models, social features tend to dominate in terms of relative feature importance.

- Most relative feature importance is found in just a small number of features, but those in the tail of the distribution also contribute, with a greater usefulness exhibited in the social random scenario.
- Using all feature classes in combination is beneficial and tends to be more highly performing with respect to my chosen metrics than just using any individual feature class.

4.3.3 Personalising classifiers

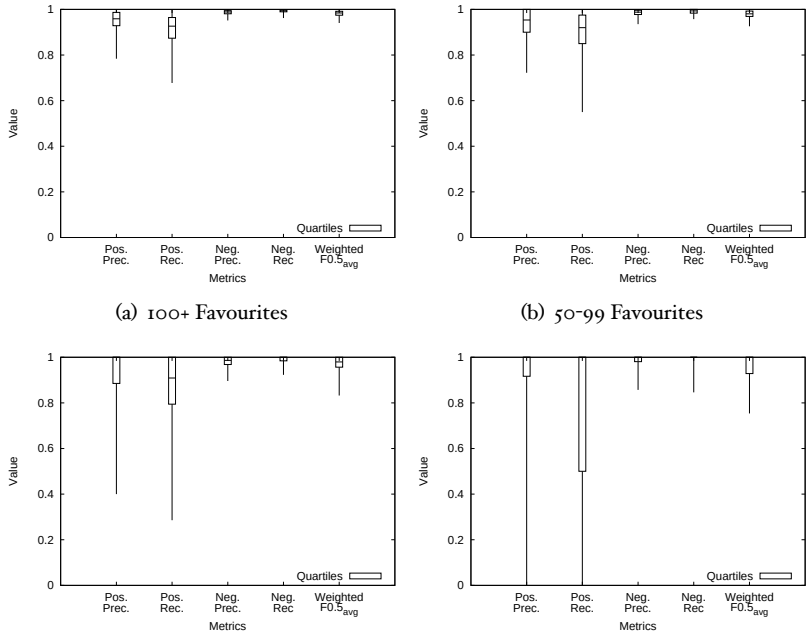
Having evaluated the models trained on four sets of users of varying labelling activity, it became clear that as performance varied between these groups, it was reasonable to assume that it would vary for users within these groups. To measure this variation in performance of my system between the 400 total individual users, the experiment was extended from doing 14 runs *per user set.* (2 data scenarios with 7 feature class combinations in each) to 2 runs *per user* (one for each data scenario, using all feature classes). The decision to do just one combination run of all feature classes for each user was based on the previous findings that the fully combined run tended to out perform the others and because it ensured total computation time was kept to a reasonable length.

4.4 Individual classifier evaluation

Figures 4.12 and 4.13 shows box and whisker plots of quartile statistics for the output performance of individually trained classifiers with respect to 5 metrics for all 400 users.

4.4.1 Across metrics

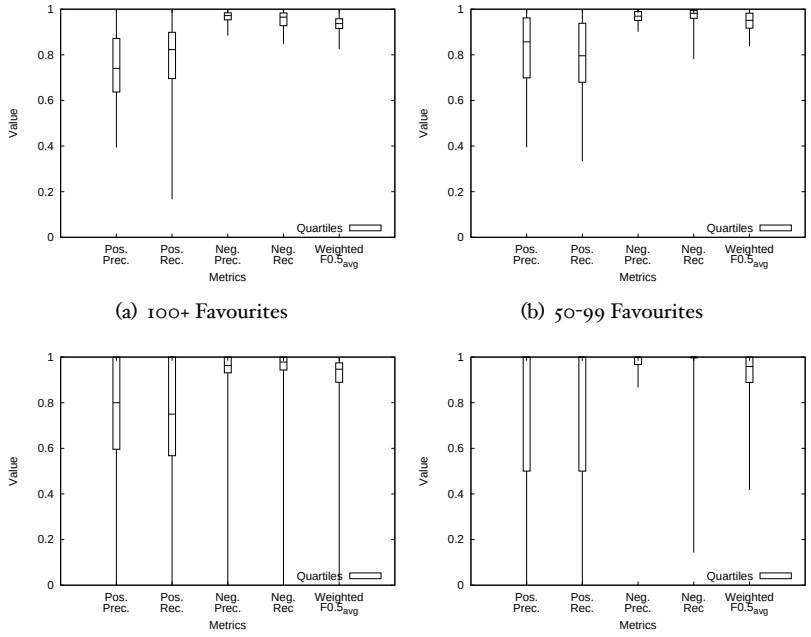
When viewed across all 8 graphs, it can be seen that metrics for both the positive and negatives examples vary, but this variation is more extensive for positive examples. This is reflected in the Weighted $F0.5_{avg}$ measure.



(a) 100+ Favourites (b) 50-99 Favourites

(c) 10-49 Favourites (d) 5-9 Favourites

FIGURE 4.12: Quartile variation of performance for individual users using the *Random* scenario.



(a) 100+ Favourites (b) 50-99 Favourites

(c) 10-49 Favourites (d) 5-9 Favourites

FIGURE 4.13: Quartile variation of performance for individual users using the *Social Random* scenario.

Looking at the positive example metrics, it can be seen that recall tends to vary more than precision, so that while the ability of the classifier to gather as many potential favourites is not very consistent, the examples it does classify as Favourites are generally correct.

The metrics for negative examples are consistently better than for the positive examples, reflecting the findings in the General Classifier approach, indicating again that the classifier is far better at accurately identifying non-Favourites than Favourites.

All metrics, in all graphs, have maximum values of 1.0. This means that in all runs, among the 100 users there is at least one instance of perfect performance for each metric. For recall, this is trivial to achieve by solely returning all possible examples, and so 1.0 is not necessarily an indicator of good performance. However, for a precision value of 1.0 to be achieved, some results have to be returned and they must all be correct. This does indicate good performance for these instances, although according to the data tables they are accompanied by poor recall.

4.4.2 Across user sets

By looking at the performance metrics between the user sets for both scenarios, we can see a relatively consistent trend in decreased consistency between results as the number of Favourites a user has decreases. This is shown by the increase in whisker length for most metrics in the graphs as Favourites per user decreases. This seems to suggest that with more Favourites to train with, a user is likely to get more consistently highly performing results from the resultant classifier. However, even for users with relatively few examples to train with, performance is still generally high, as shown by the median bars in the graph and the weighted $F0.5_{avg}$ measure.

It can also be observed that performance values become more polarised as the Favourites per user decreases, in that by Figure 4.12(d) and Figure 4.13(d), many individual values are either 1.0 or 0.0, contributing to average metrics typified by the positive recall in both those

subgraphs. Of course, with fewer examples to judge, the results are likely to be more granular anyway, but for the 5-9 Favourites user group there are sufficient examples to expect more granular metrics. It could therefore be the case that at this level of user Favourite labelling, the classifier either performs well, or not at all (unlike the more graduated performance for the other user sets), hence leading to these almost binary performance metrics.

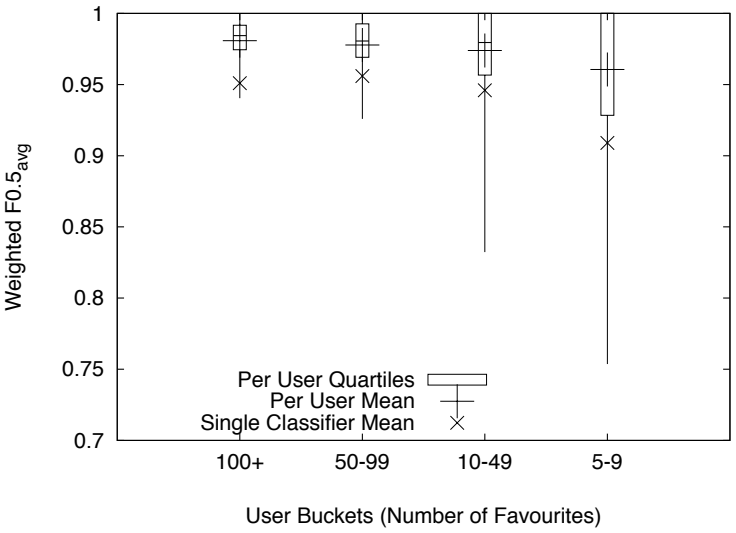
4.4.3 Between scenarios

When comparing between the two data scenarios, the most striking variation is how much worse and inconsistent the *Social Random* scenario results are relative to the *Random* scenario. This reflects the findings of the General Classifier approach, but with this per user analysis it can be seen that it is the inconsistency that is dragging down overall performance metrics and accounts or the variation between the two scenarios. This suggests that by selecting negative examples using social criteria, the resultant classifier behaves more erratically with respect to my metrics. It also shows how much of an influence the choice in negative training examples has on the performance on positive examples.

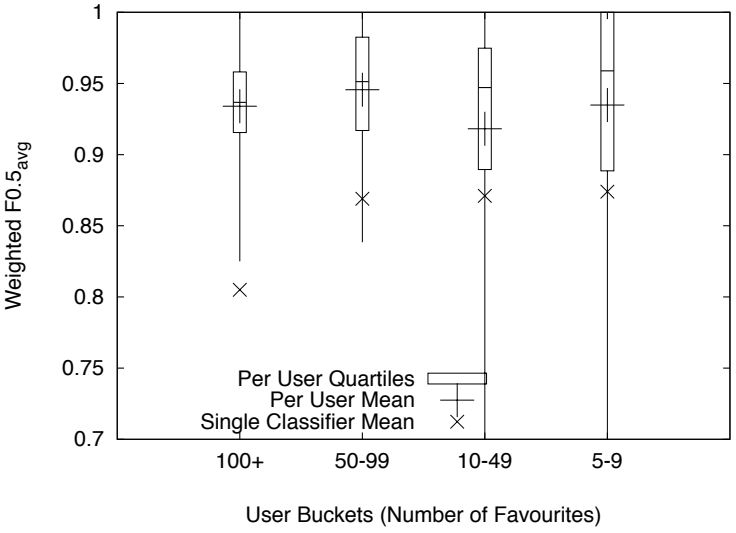
4.4.4 Comparison with general classifier approach

In order to compare the individually trained classifiers and the single general classifier approach, I focus on the Weighted $F0.5_{avg}$ measure that sums up performance for both precision and recall.

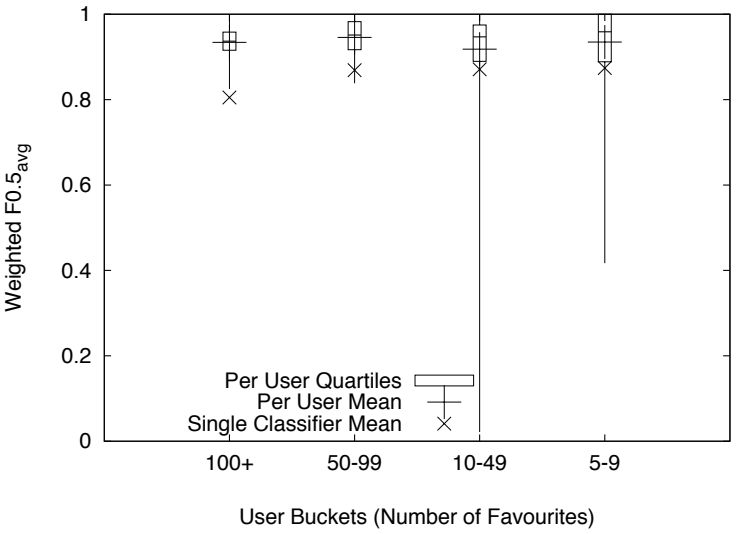
Figure 4.14 shows the Weighted $F0.5_{avg}$ measure for each of the four user sets in addition to the box and whisker plot of the quartile statistics from the performance of the individually trained classifiers. For the Random Scenario, I show the graph (Figure 4.14(a)) with x-axis range $0.7 < F0.5_{avg} < 1.0$. For the Social Random scenario, I show the full graph (Figure 4.14(c)) with x-axis range $0.0 < F0.5_{avg} < 1.0$ as well as $0.7 < F0.5_{avg} < 1.0$ (Figure 4.14(b)) to allow easier inspection of values near the upper end of the range.



(a) *Random* Scenario ($0.7 <$ x-axis range < 1.0)



(b) *Social Random* Scenario ($0.7 <$ x-axis range < 1.0)



(c) *Social Random* Scenario ($0.0 <$ x-axis range < 1.0)

FIGURE 4.14: The comparison between the performance statistics of the individually trained classifiers and the single general classifier performance.

It can be seen that the general classifier results are consistently below the mean values as calculated over all the individual classifiers for that user bucket, showing that on average, most users would get better performance with their own personalised classifier. However, for all user buckets except that in the *Social Random* scenario with 100+ Favourites, there are users for whom the general classifier outperforms the individually trained ones.

This implies that there is no single best approach for all users with respect to their number of Favourites.

4.4.5 Summary of individually trained model findings

- The more favourites a user has, the higher and more consistently well performing the resultant classifier is.
- There is a wide variation in performance between users, seemingly correlated with the favourite labelling activity of user.
- Negative examples are still easier to predict than positive class (similar to general approach).
- Individually trained trees offer potential for marginally highly performance, but at the cost of increased computation.
- There are users for whom the general classifier performs better.

4.4.6 The value of personalising classifiers

The previous section showed that neither the general nor the individual classifier approach outperforms the other consistently. While the individually trained classifiers have the potential to be more highly performing for most users, for those with very little training data (i.e. those with few Favourites), the general classifier can be better. Determining exactly which users this is the case for, I leave to future work.

In the cases where individually trained classifiers is the best performing option, the cost of this approach must be taken in to account. While it may be feasible to train 400 classifiers as was done in this experiment, scaling this up to the millions of users in a system like Flickr or Facebook would entail a linear growth in computation. This also ignores the cost of retraining that may be required when users label more Favourites that can be used for increasing training data. As an example, for users with 100+ Favourites, a GBDT model can be trained for a single user in on average 70 seconds on a single standard 2.5GHz dual core desktop computer. For users with 5-9 Favourites, this time was reduced to 16s. However, for the 3.2 millions users who I estimate to have 100+ Favourites (according to the distribution in Figure 4.2), and assuming my average timings are realistic, this would require over 7 years worth of single computer computation time for these users alone. Or on a cluster of 200 similar machines, this could be reduced to around 13 days.

A single classifier, however, can be trained once for all users, and only that single classifier would ever need retraining.

The scaling and computation issues of individual classifiers could be mitigated by finding a compromise approach, whereby users are clustered into groups of similar users, and a classifier trained not for each user, but for each cluster. This would reduce the total number classifiers required while maintaining some of the greater performance of the more personalised approach.

4.5 Conclusions

This chapter started with the task of trying to predict, for a particular user, photos from an incoming stream of previously unseen images that the user would be likely to label a favourite, and analyse the contribution of social features when solving this problem.

By modelling the task as a classification problem, I have used machine learning techniques trained on features extracted from the image, its metadata and it social context to provide

high performance judgements for users who have varying quantities of previously labelled favourite images.

In implementing this system I have shown that accurately detecting pseudo-negative examples (non-Favourites) is far easier and more consistent than detecting positive examples (Favourites). I have shown this to be true regardless of the two data scenarios I devised to approximate unavailable negative feedback examples, and regardless of the level favourite labelling activity of the user.

Being able to so effectively detect negative examples is in itself a useful outcome as this can be used to prune out irrelevant images for users when searching and browsing, reducing the mental burden on the user and speeding up their interactions as well as potentially reducing computation costs for online media sharing systems.

I have analysed the performance of my system not only for active users but for a number of sets of users, as well as individually and shown how performance varies with respect to labelling activity. By doing so I have shown what levels of performance a service provider can expect from such a system and provided baseline performance measures for future work to be based on.

While it was not possible to test my system on users who had no previously favourite labelling activity within the confines of the experiment and the resources available, I am confident the general classifier trained for multiple users would still produce useful results.

I have also shown that training an individual classifier for each user offers the potential for slightly higher overall performance, but with the cost of high computation costs. A system that melded the general classifier with individually trained ones, depending on the attributes of the users, could provide the best of both worlds—the high performance of the individually trained classifiers with the coverage of the general classifier.

I have analysed the three features classes (Social, Textual and Visual) with respect to their contribution to the performance of my classifiers, both as individually and in combination

with each other. I have shown the social features dominate those which are most important to the GBDT classifiers trained, as well as how content based features can be useful to augment the more traditional textual approaches found in current literature.

In particular, this chapter has addressed three of my initial hypothesis sub-questions from Section 1.2.2:

Which social connections yield the most valuable information for use in tag and image recommender systems designed for large online photo sharing systems? I have shown in Tables4.8, 4.7, 4.7 and 4.5 the relative importance of the top 10 most important features for each of the four user sets and for both scenarios. In Table 4.9 I summarise these results further by showing the top three features throughout all runs for both scenarios. Social feature are highly discriminating with respect to classifying Favourites, more so in the *Social Random*~ scenario than in the *Random*~ scenario.

How can social connections be most effectively used to improve recommendation? The machine learnt classifier presented in this chapter has been able to use features which describe the social connection between a viewer and an owner of an image to make accurate judgements over my test data. In runs not using social features, and in what I believe is a more realistic dataset (the *Social Random*~ scenario) $F0.5_{avg}$ varies between 0.530 and 0.682. However, when social features are added, they were able to boost performance with respect to $F0.5_{avg}$ to between 0.805 to 0.874.

In comparison with the other two feature classes, it was the social features that were most important when training.

How can different kinds (textual/visual/social) of media/user descriptors be combined effectively in an image recommender system? I have shown that the textual metadata associated with an image, in addition to features extracted from visual content

can be effectively combined with data describing the social context of the user to provide effective image recommendations. This combination, whilst dominated by social features, also has significant contribution from the other two classes. The combination of all three classes was shown to usually outperform any of the individual runs based on just one feature class.

Ultimately, by using the system as that proposed, implemented and evaluated in this chapter I can reduce the number of irrelevant images shown to a user, boost the number of relevant ones, all automatically, with very good performance for most users. The high performance is mostly due to the extraction of effective features that describe social context.

4.6 Reflection on Flickr Favourites

This chapter has presented a technique that can be used to extend existing media handling systems by tailoring the experience to specific users in a way that no current system provides for. It can be used to present images to users that they are likely to find more relevant than non-personalised retrieved images and I have shown how the benefit of this approaches changes depending on the previous Favourite labelling of users.

These findings can be integrated into existing systems, particularly into those that have gathered forms of social context data but have not yet exploited it in the manner I have presented. This would include Flickr, on which I based my experiments, as well as similar systems like Picasa Web Albums, Facebook and Photobucket (amongst a growing number of others).

Between these potential venue for exploitation of my findings, they could have an impact of hundreds of millions of web users, increasing their satisfaction within the use-cases of image searching and browsing in terms of relevance of returned images and the speed with which they are fetched (instead of returning hundreds of potentially relevant images, a sub set of highly probably relevant images can be returned instead).

I have shown how the particular contribution of using the social context of users in image recommendation can be effectively augmented by using it in concert with textual and visual information. This combination of sources for this specific task is also not used in existing systems. By quantifying the value of using such an approach over reasonable baselines, I have provided evidence that such information combination is worth further research and development by systems that could be improved by it.

In addition to the task of image classification presented in this chapter, my approach could be used to improve other online, non-image sharing systems. The technique could be easily extended to intuitively similar video and audio sharing systems that incorporate an element of social activity between users. It would, therefore, not be appropriate for 'broadcast' systems like the BBC iPlayer[12] in which media is consumed by users, but in which they have very little interaction between themselves.

Other online systems that previously did not have social elements (news and shopping sites for example) are increasingly incorporating mechanisms for their users to interact. By using the characterisation of user interaction I have shown in this chapter (tailored, of course, to their own users and media-equivalents) I suggest they will also be able to improve user satisfaction and engagement.

[12] http://www.bbc.co.uk/iplayer/

Chapter 5

Conclusions and Future Direction

"While I'm still confused and uncertain, it's on a much higher plane, d'you see, and at least I know I'm bewildered about the really fundamental and important facts of the universe."

Treatle nodded. "I hadn't looked at it like that," he said.

"But you're absolutely right. He's really pushed back the boundaries of ignorance."

Discworld scientists at work

(Terry Pratchett, Equal Rites)

Roadmap The previous chapters have dealt with introducing my hypothesis, contextualising it with respect to the current state-of-the-art and presenting experimental work that was implemented to prove my hypothesis true and answer its sub-questions.

This chapter evaluates my approach to testing my hypothesis and analyses my findings. Each of the hypothesis sub-questions is addressed with respect to the work presented in previous chapters. I then discuss the potential for further investigation.

5.1 Hypothesis evaluation

In order to judge whether I have adequately tested my hypothesis I first address the hypothesis sub-questions presented in Section 1.2.2, with respect to my two complementary sets of experiments.

In Chapter 3 I presented the problem of recommending tags to a user as they are annotating their images and do so within the context of the online image sharing website Flickr. I showed both in my related work evaluation in Chapter 2 how techniques have been used to tackle this problem in the past, including using collective tag co-occurrences from throughout an image dataset, as well as using tag co-occurrences generated from a user's own tag vocabulary.

In order to improve upon these techniques, I introduced a graph-based formalisation of these tag co-occurrences that allows any set of photos and their tags to be represented in a consistent and well-defined manner. I use this in combination with a probabilistic framework to make tag suggestions.

In Chapter 4 I introduce the task of image recommendation within the context of predicting which images in an incoming stream of new image a user is likely to label as a Favourite in Flickr. I treat this task as a binary classification problem, deciding between whether an image would be labelled a Favourite or not. In order to tackle this problem I propose a machine learnt approach that extract textual, visual and social features from a new image and uses those features to make a judgement.

Which social connections yield the most valuable information for use in tag and image recommender systems designed for large online photo sharing systems? With respect to tag recommendation, the evaluation of my tag suggestion approach shows that the graph based on the 'Contact' relationship did not usefully contribute tag co-occurrence information for tag suggestion with respect to my chosen metrics.

The group graph was also no better than the baseline graph when evaluated by itself. However, there was a significant increase in performance when it was combined with the collective with respect to two of my metrics when evaluated over all the users in my dataset.

While the most explicit connection users can make in Flickr (Contacts) was found not to be useful when making tag suggestions, the less explicit group membership connection was.

With respect to image recommendation, I showed that in my approach to predict Favourite images, between the three classes of features, the implemented social features exhibited greater relative importance than the other two classes, variance depending on the number of Favourite images the viewer had. Of those social features, whether the uploader of an image was a contact of the viewing user was consistently important between my two datasets, contrasting with the value exhibited by this social relationship in the tag suggestion experiment. The number of times the image had been viewed by other users was also a valuable feature, as was the number of contacts the viewer and uploader shared in common and whether the viewed image was in a group of which the viewer was a member.

How can these social connections be effectively used to improve recommendation? With respect to tag recommendation, the social graph could not be used effectively to improve performance. It consistently performed worse than the baseline approach based on collective data when compared in isolation and evaluated over all users in my dataset, as well as when users were split according to the number of contacts they had. When the tag suggestions provided by this data were combined with the baseline collective graph, the personal graph and the group graph, the contact graph was never significantly beneficial and usually retarded performance.

However, I showed how the group graph significantly increased performance when combined with both the collective graph and the personal graph when evaluated over all users. When performance was broken down according to the number of groups of which a user

was a member, in most cases performance was increased, but no correlation was found with group membership.

With respect to image recommendation, I demonstrated that among the basket of commonly used textual and visual features, the social features frequently exhibited the highest relative importance when training the gradient boosted decision trees used in my experiment. When runs were evaluated based on the individual classes of features (textual, visual and social), the social features always had a higher $F0.5_{avg}$ (my principle evaluation metric), regardless of the negative example dataset used and regardless of the number of Favourites the user in question had labelled.

In comparison to the runs involving just textual and visual features, when social features were added, performance with respect to $F0.5_{avg}$ was always higher, again demonstrating the value of these features for this task (in the context of the exact features used and the dataset, etc.).

How can different kinds (textual/visual/social) of media/user descriptors be combined effectively in a image recommender system? In Chapter 4 I chose the state-of-the-art gradient boosted decision tree technique for learning to classifier images between Favourite and non-Favourite. This approach allowed me to effectively combine features derived from all three classes to produce a single highly performing classifier, as evidenced by the results that showed that the full combination of all features usually led to a better classifier than any one individual class with respect to my metrics and usually better than any other combination of classes of features.

Can single positive feedback cues like the Flickr Favourite label be used to train systems to predict further Favourites? In Chapter 4 I use the Favourite label as an indicator of positive approval for a particular image given by a specific user. I showed how by using a combination of three feature classes I could train a highly discriminating classifier.

When applied to previously unseen data, this classifier was able to effectively judge images as likely to be Favourites or not.

I also demonstrated this approach with respect to two variant negative sample datasets and analysed how their construction effected output classifier performance, and in both cases the Favourite label used as positive data was still able to train a high performing classifier.

Hypothesis evaluation To reiterate:

Existing techniques for managing large-scale online image collections do not currently fully take advantage of the rich social context of the data itself and the users who interact with it, a form of data that is increasingly available. Nor do they leverage the social connections between people who use such systems. By accurately modelling these connections and understanding more about them, we learn more about user image and tag preference. **More specifically, image and related metadata recommender systems can be built using this social information that are more effective than existing state-of-the-art non-social techniques.**

Using two sample use-cases of recommendation in photo system interaction based in Flickr, I have shown that tag and image recommendation can be improved over baselines that did not use social information that I chose based on their prevalence and standing in the research community. However, not all social connections were found to be equally useful between the two example recommendation tasks. For example, I showed that while contacts are not useful when suggesting tags, they are effective for predicting favourite images, showing that the specific value of social features is tied to the recommendation task to which they are applied.

With the increasing availability of social data associated with media like photos, the exploitation of this extra source of information will become even more valuable. Tailoring the experience of individual users using personalised recommendation approaches as demonstrated in this thesis will ensure users get to the content they want quicker and more efficiently than before, increasing their satisfaction with their interaction with such systems.

5.2 Limitations and future work

Discovering implicit user clusters according to social behaviour During the work in this thesis, I have shown which of the social features I presented were useful for specific recommendation use-cases in large online web sharing environments. During my evaluation of results, I looked at how performance varied between users and between demographic buckets of users. This grouping of users led to greater insight into how performance varied among a large, varied community of users.

This analysis could be extended to look at whether the kind of social features I have presented here (or others) be used to discover new, previously unrecognised implicit communities within Flickr. These may be based on shared behaviour or attributes that users may not be aware of and could, like the social context used in this thesis, be exploited for improved image recommendation. Perhaps more interestingly from a sociological perspective, the correlation between such communities and real-world social connections could also be measured.

These avenues of research would help to increase understanding of online communities that form in online media sharing environments and how they relate to their offline social counterparts.

Aesthetics of images In Chapter 4 I trained classifiers to find Favourite images in Flickr. I used the Favourite label as an indicator of a users interest in images and as a sign of their approval. These images could be analysed further to improve existing photographic guidelines for what constitutes 'good' photos, or those that are suited for specific purposes. This be further broken down to find guidelines for certain genres of photographs (portraiture, landscapes, urban, etc.) by focusing on the feedback photo sharing communities give to photo in those styles.

Existing guidelines for taking photos tend to be based purely on visual rules—composition, use of colour, lighting, etc. Could these be extended with greater understanding of the semantic and social characterisations of images, automatically derived from images given labels like Flickr's Favourite label?

Using task marketplaces to evaluate personalised recommendation experiments

The evaluation techniques for the recommendation systems presented in this thesis have been automated methods. They were chosen to avoid the difficult and expensive process of evaluating the systems' output with real users. While my approaches were well suited to the experiments they were used in, the value of more direct user feedback is significant.

New human interaction task marketplace systems like Amazon's Mechanical Turk[1], CloudCrowd[2] and Clickworker[3] provide a task publisher with a forum to advertise a short simple online task (e.g. evaluating a recommendation system's output) and a framework to actually allow the task to take place. These systems allow for the large scale distribution of tasks with some control over target demographics, opening up new opportunities for evaluation using larger numbers of real users compared to what is usually feasible in an academic research environment. The costs involved also make it viable for researchers to collect more data in this way than through traditional methods.

Personalisation is difficult to introduce in what is otherwise an anonymous working environment, but I have already started investigations into how this can be achieved (although outside the scope of this thesis).

Figure 5.1 shows two screenshots of an interface I call *Predictr* that gathers feedback regarding the quality of personalised recommendations. Figure 5.1(a) asks a participant, for whom a personalised classifier has been trained as they signed up for the task, to pick the set of images they would judge to contain more images they would label as favourites. Figure 5.1(b)

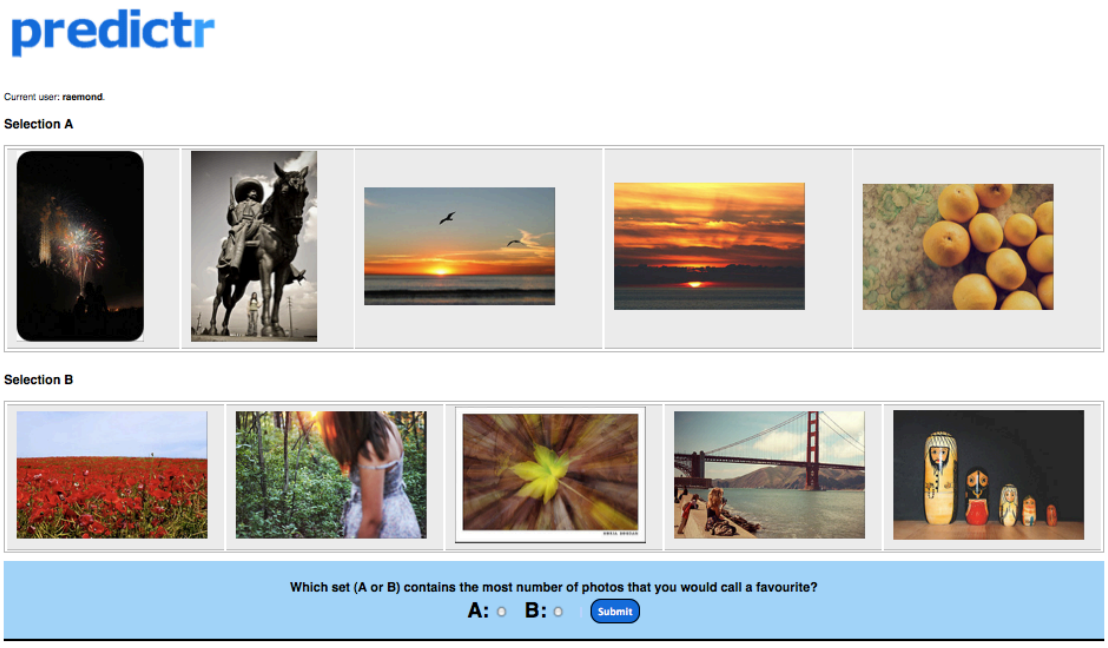
[1] http://www.mturk.com/mturk/
[2] http://www.cloudcrowd.com/
[3] http://www.clickworker.com/

asks for more specific comparative judgements that would allow for the inducement of a linear scale of preference calculated over repeated tasks given to the user.

These two interface have both been tested and form the basis for continuing experiments.

The introduction to this thesis highlighted the recent growth of online media, how much the phenomenon has developed from being used by a small number of technically minded users to the general public, as well as how it has diversified. From the first cumbersome experimental electronic cameras being used by a few scientists to Apple's iPhone 4 being the most popular device for taking photographs on Flickr[4]. Users are generating and consuming media faster than ever before and the systems designed to handle such data have to evolve with the changing requirements of their growing, ever more social communities. The research I present in this thesis contributes in a small way to supporting these users.

[4] http://www.flickr.com/cameras/ as of August 2011, http://techcrunch.com/2011/06/21/flickr-iphone-data/



(a) Version for comparing two sets of images, one of which has been chosen as likely Favourites, the other as non-Favourites.

(b) Version for comparative evaluation of single images.

FIGURE 5.1: The *Predictr* interface for manual evaluation of image recommendation for use with Mechanical Turk.

Appendix

TABLE 5.1: The results from the General Classifier approach experiment as presented in Section 4.3.1, for users with 100+ favourites as specified in Section 4.1.2, with textual (T), visual (V) and social (S) features. Trivial classifier shown for comparison (see Section 4.2.3).

Features	Favourite		non-Favourite		$F0.5_{avg}$
	Precision	Recall	Precision	Recall	
Random Scenario					
T	0.845	0.480	0.925	0.986	0.760
V	0.781	0.148	0.883	**0.994**	0.486
S	0.939	0.821	0.973	0.992	0.921
T+V	0.875	0.575	0.938	0.987	0.813
T+S	0.950	0.827	0.974	0.993	0.930
V+S	0.945	0.907	**0.986**	0.992	0.944
T+V+S	**0.954**	**0.910**	**0.986**	0.993	**0.951**
Social Random Scenario					
T	0.469	0.174	0.884	0.970	0.423
V	**1.000**	0.085	0.877	**1.000**	0.394
S	0.803	0.560	0.935	0.979	0.766
T+V	0.614	0.269	0.896	0.974	0.545
T+S	0.771	**0.620**	0.943	0.972	0.764
V+S	0.886	0.560	0.936	0.989	**0.814**
T+V+S	0.838	0.619	**0.944**	0.982	0.805
Trivial Classifier	0.133	0.133	0.866	0.866	0.769

Table 5.2: The results from the General Classifier approach experiment as presented in Section 4.3.1, for users with 50-99 favourites as specified in Section 4.1.2, with textual (T), visual (V) and social (S) features. Trivial classifier shown for comparison (see Section 4.2.3).

Features	Favourite		non-Favourite		$F0.5_{avg}$
	Precision	Recall	Precision	Recall	
Random Scenario					
T	0.927	0.594	0.941	0.993	0.849
V	0.713	0.146	0.883	0.991	0.469
S	0.934	0.783	0.968	0.991	0.909
T+V	0.944	0.667	0.951	0.994	0.883
T+S	**0.971**	0.815	0.972	**0.996**	0.941
V+S	0.939	0.866	0.980	0.991	0.931
T+V+S	0.969	**0.889**	**0.983**	**0.996**	**0.956**
Social Random Scenario					
T	0.478	0.220	0.890	0.963	0.456
V	0.795	0.084	0.877	**0.997**	0.375
S	0.896	0.607	0.943	0.989	0.836
T+V	0.619	0.289	0.899	0.973	0.559
T+S	0.881	0.649	0.948	0.987	0.840
V+S	**0.913**	0.672	0.952	0.990	0.866
T+V+S	0.896	**0.721**	**0.959**	0.987	**0.869**
Trivial Classifier	0.133	0.133	0.866	0.866	0.769

Table 5.3: The results from the General Classifier approach experiment as presented in Section 4.3.1, for users with 10-49 favourites as specified in Section 4.1.2, with textual (T), visual (V) and social (S) features. Trivial classifier shown for comparison (see Section 4.2.3).

Features	Favourite		non-Favourite		$F0.5_{avg}$
	Precision	Recall	Precision	Recall	
Random Scenario					
T	0.906	0.519	0.931	0.992	0.809
V	0.641	0.139	0.882	0.988	0.443
S	0.906	0.801	0.970	0.987	0.895
T+V	0.910	0.561	0.937	0.992	0.828
T+S	0.949	0.827	0.974	0.993	0.929
V+S	0.930	0.867	0.980	0.990	0.925
T+V+S	**0.955**	**0.885**	**0.983**	**0.994**	**0.946**
Social Random Scenario					
T	0.465	0.218	0.890	0.962	0.449
V	0.540	0.070	0.875	**0.991**	0.320
S	0.876	0.645	0.948	0.986	0.836
T+V	0.607	0.249	0.895	0.976	0.530
T+S	0.907	0.642	0.948	0.990	0.854
V+S	0.904	0.660	0.950	0.989	0.857
T+V+S	**0.917**	**0.679**	**0.953**	**0.991**	**0.871**
Trivial Classifier	0.133	0.133	0.866	0.866	0.769

TABLE 5.4: The results from the General Classifier approach experiment as presented in Section 4.3.1, for users with 5-9 favourites as specified in Section 4.1.2, with textual (T), visual (V) and social (S) features. Trivial classifier shown for comparison (see Section 4.2.3).

Features	Favourite		non-Favourite		$F0.5_{avg}$
	Precision	Recall	Precision	Recall	
Random Scenario					
T	0.765	0.474	0.991	0.997	0.723
V	0.511	0.108	0.985	0.998	0.385
S	0.916	**0.822**	**0.997**	**0.999**	**0.909**
T+V	0.839	0.465	0.991	0.998	0.759
T+S	0.902	0.817	**0.997**	0.998	0.898
V+S	0.901	0.808	**0.997**	0.998	0.896
T+V+S	**0.920**	0.808	**0.997**	**0.999**	**0.909**
Social Random Scenario					
T	0.521	0.408	0.990	0.993	0.560
V	0.208	0.023	0.983	0.998	0.202
S	0.832	0.718	0.995	0.997	0.831
T+V	0.733	0.413	0.990	0.997	0.682
T+S	0.890	**0.756**	**0.996**	0.998	**0.877**
V+S	**0.910**	0.662	0.994	**0.999**	0.866
T+V+S	0.904	0.704	0.995	**0.999**	0.874
Trivial Classifier	0.133	0.133	0.866	0.866	0.769

Bibliography

Gediminas Adomavicius and Alexander Tuzhilin, 2005. Toward the next generation of recommender systems: A survey of the State-of-the-Art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749. ISSN 1041-4347.

Morgan Ames and Mor Naaman, 2007. Why we tag: motivations for annotation in mobile and online media. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 971–980, San Jose, California, USA. ACM. ISBN 978-1-59593-593-9.

Chris Anderson, 2008. *The Long Tail: Why the Future of Business Is Selling Less of More*. Hyperion. ISBN 9781401309664.

Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan, August 2006. Group formation in large social networks: Membership, growth and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, Philadelphia, Pennsylvania, USA.

Brian Bartell, Garrison Cottrell, and Richard Belew, 1994. Automatic combination of multiple ranked retrieval systems. *Proceedings of the 17th Annual International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 173–181.

Herbert Bay, Tinne Tuytelaars, and Luc Van Gool, May 2006. SURF: speeded up robust features. *Proceedings of the Ninth European Conference on Computer Vision*.

Jon Louis Bentley, September 1975. Multidimensional binary search trees used for associative searching. *Commun. ACM*, 18:509–517. ISSN 0001-0782. URL http://doi.acm.org/10.1145/361002.361007.

Matthew Boutell and Jiebo Luo, June 2005. Beyond pixels: Exploiting camera metadata for photo classification. *Pattern Recognition*, 38(6):935–946. ISSN 0031-3203.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender, 2005. Learning to rank using gradient descent. *Proceedings of the 22nd international conference on Machine learning*, page 89–96. ACM ID: 1102363.

Iván Cantador Gutiérrez, October 2008. *Exploiting the conceptual space in hybrid recommender systems: a semantic-based approach*. PhD thesis, Universidad Autonoma de Madrid, Madrid.

Rich Caruana and Alexandru Niculescu-Mizil, 2006. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, page 161–168, Pittsburgh, Pennsylvania. ACM. ISBN 1-59593-383-2. ACM ID: 1143865.

Meeyoung Cha, Alan Mislove, and Krishna P. Gummadi, 2009. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th international conference on World wide web*, pages 721–730, Madrid, Spain. ACM. ISBN 978-1-60558-487-4.

D. Chai, S. L. Phung, and A. Bouzerdoum, 2003. A bayesian skin/non-skin colour classifier using non-parametric density estimation. *Proceedings of the 2003 International Symposium on Circuits and Systems*, 2:464–467.

Heather Champ. 4,000,000,000 « flickr blog. http://blog.flickr.net/en/2009/10/12/4000000000/, October 2009.

Savvas Chatzichristofis and Yiannis Boutalis, 2008. CEDD: color and edge directivity descriptor: A compact descriptor for image indexing and retrieval. In *Proceedings of the 6th international conference on Computer vision systems*, pages 312–322. Springer-Verlag.

Ed H. Chi and Todd Mytkowicz, 2008. Understanding the efficiency of social tagging systems using information theory. In *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 81–88, Pittsburgh, PA, USA. ACM. ISBN 978-1-59593-985-2.

Ritendra Datta, Dhiraj Joshi, Jia Li, and James Wang, 2006. Studying aesthetics in photographic images using a computational approach. In *Computer Vision–*, pages 288–301.

M. De Choudhury, H. Sundaram, A. John, and D. D. Seligmann, 2009a. Social synchrony: Predicting mimicry of user actions in online social media. In *Computational Science and Engineering, 2009. CSE'09. International Conference on–*, volume 4, page 151–158.

M. De Choudhury, H. Sundaram, A. John, and D. D Seligmann, 2009b. What makes conversations interesting?: themes, participants and consequences of conversations in online social media. In *Proceedings of the 18th international conference on World wide web*, page 331–340.

M. De Choudhury, H. Sundaram, Yu-Ru Lin, A. John, and D.D. Seligmann, 2009c. Connecting content to community in social media via image content, user tags and user communication. In *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on–*, pages 1238–1241. ISBN 1945-7871.

Jeffrey Dean and Sanjay Ghemawat, 2008. MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113.

Reinhard Diestel, February 2006. *Graph Theory (Graduate Texts in Mathematics)*. Springer. ISBN 3540261834.

P. W Eklund and A. Hoang, 2002. A performance survey of public domain supervised machine learning algorithms. *Australian Journal of Intelligent Information Systems. v9 i1*, page 1–47.

Charles Elkan and Keith Noto, 2008. Learning classifiers from only positive and unlabeled data. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 213–220, Las Vegas, Nevada, USA. ACM. ISBN 978-1-60558-193-4.

Facebook. 10 billion photos. http://www.facebook.com/note.php?note_id=30695603919, October 2008.

Facebook. Facebook statistics. http://www.facebook.com/press/info.php?statistics, October 2010.

Flickr. Flickr: The help forum: [Closed] how many flickr' users?!? http://www.flickr.com/help/forum/en-us/97258/, May 2009.

D. Frank Hsu and Isak Taksa, January 2005. Comparing rank and score combination methods for data fusion in information retrieval. *Information Retrieval*, 8(3):449–480. ISSN 1386-4564.

Jerome H. Friedman, October 2001. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232. ISSN 00905364.

Jing Gao, Wei Fan, Yizhou Sun, and Jiawei Han, 2009. Heterogeneous source consensus learning via decision propagation and negotiation. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, page 339–348, New York, NY, USA. ACM. ISBN 978-1-60558-495-9. ACM ID: 1557061.

Nikhil Garg and Ingmar Weber, October 2008. Personalized, interactive tag recommendation for flickr. In *Proceedings of the 2008 ACM Conference on Recommender Systems*, pages 67–74, Lausanne, Switzerland. ACM. ISBN 978-1-60558-093-7.

Rafael C. González and Richard Eugene Woods, 2008. *Digital image processing*. Prentice Hall. ISBN 9780131687288.

Jefferson Graham, February 2006. Flickr of idea on a gaming project led to photo website. *USA TODAY*.

M. Grubinger, P. Clough, H. Müller, and T. Deselaers, 2006. The IAPR TC-12 benchmark-a new evaluation resource for visual information systems. In *International Workshop OntoImage*, page 13–23.

David Hasler and Sabine E. Suesstrunk, June 2003. Measuring colorfulness in natural images. In Bernice E. Rogowitz and Thrasyvoulos N. Pappas, editors, *Human Vision and Electronic Imaging VIII*, volume 5007, pages 87–95, Santa Clara, CA, USA. SPIE.

Rui Hu, S. Ruger, Dawei Song, Haiming Liu, and Zi Huang, 2008. Dissimilarity measures for content-based image retrieval. In *Multimedia and Expo, 2008 IEEE International Conference on*, pages 1365–1368.

Jeff Huang, Katherine M Thornton, and Efthimis N Efthimiadis, 2010. Conversational tagging in twitter. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, HT '10, page 173–178, New York, NY, USA. ACM. ISBN 978-1-4503-0041-4. ACM ID: 1810647.

Kai-Qi Huang, Qiao Wang, and Zhen-Yang Wu, July 2006. Natural color image enhancement and evaluation algorithm based on human visual system. *Computer Vision and Image Understanding*, 103(1):52–63. ISSN 1077-3142.

M. J Huiskes and M. S Lew, 2008a. The MIR flickr retrieval evaluation. In *Proceeding of the 1st ACM international conference on Multimedia information retrieval*, page 39–43.

M. J Huiskes and M. S Lew, 2008b. Performance evaluation of relevance feedback methods. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*, page 239–248.

Mark J. Huiskes, Bart Thomee, and Michael S. Lew, 2010. New trends and ideas in visual concept detection: The MIR flickr retrieval evaluation initiative. In *MIR '10: Proceedings*

of the 2010 ACM International Conference on Multimedia Information Retrieval, pages 527–536, New York, NY, USA. ACM.

Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich, 2010. *Recommender Systems: An Introduction*. Cambridge University Press. ISBN 9780521493369.

Robert Jäschke, Leandro Marinho, Andreas Hotho, Lars Schmidt-Thieme, and Gerd Stumme, 2007. Tag recommendations in folksonomies. In *Proceedings of the 11th European conference on Principles and Practice of Knowledge Discovery in Databases*, volume 4702, pages 506–514, Warsaw, Poland. Springer-verlag. ISBN 978-3-540-74975-2.

P. Kakumanu, S. Makrogiannis, and N. Bourbakis, March 2007. A survey of skin-color modeling and detection methods. *Pattern Recognition*, 40(3):1106–1122. ISSN 0031-3203.

Yan Ke and Rahul Sukthankar, 2004. PCA-SIFT: a more distinctive representation for local image descriptors. In *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, volume 2, pages 506–513, Los Alamitos, CA, USA. IEEE Computer Society.

Yan Ke, Xiaoou Tang, and Feng Jing, 2006. The design of High-Level features for photo quality assessment. In *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, volume 1, pages 419–426, Los Alamitos, CA, USA. IEEE Computer Society.

R. Kern, M. Granitzer, and V. Pammer, May 2008. Extending folksonomies for image tagging. In *Image Analysis for Multimedia Interactive Services, 2008. WIAMIS '08. Ninth International Workshop on*, pages 126–129.

R. D King, C. Feng, and A. Sutherland, 1995. Statlog: comparison of classification algorithms on large real-world problems. *Applied Artificial Intelligence*, 9(3):289–333.

Ioannis Konstas, Vassilios Stathopoulos, and Joemon M. Jose, 2009. On social networks and collaborative recommendation. In *Proceedings of the 32nd international ACM SIGIR*

conference on Research and development in information retrieval, pages 195–202, Boston, MA, USA. ACM. ISBN 978-1-60558-483-6.

S. B. Kotsiantis, 2007. Supervised machine learning: A review of classification techniques. In *Proceeding of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, page 3–24.

Ravi Kumar, Jasmine Novak, and Andrew Tomkins, 2006. Structure and evolution of online social networks. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 611–617, Philadelphia, PA, USA. ACM. ISBN 1-59593-339-5.

Eugene F. Lally, 1961. Mosaic guidance for interplanetary travel. In *Space Flight Report to the Nation*.

Joon Ho Lee, 1997. Analyses of multiple evidence combination. In *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 267–276, Philadelphia, Pennsylvania, United States. ACM. ISBN 0-89791-836-3.

Kristina Lerman, December 2006. Social networks and social information filtering on digg. *Arxiv preprint cs/0612046*.

Kristina Lerman and Laurie Jones, December 2006. Social browsing on flickr. In *Proceedings of International Conference on Weblogs and Social Media*.

David Liben-Nowell, Jasmine Novak, Ravi Kumar, Prabhakar Raghaven, and Andrew Tomkins, August 2005. Geographic routing in social networks. *PNAS*, 102(33):11623–11628.

Tjen-Sien Lim, Wei-Yin Loh, and Yu-Shan Shih, 2000. A comparison of prediction accuracy, complexity, and training time of Thirty-Three old and new classification algorithms. *Machine Learning*, 40(3):203–228. ISSN 0885-6125. 10.1023/A:1007608224229.

G. Linden, B. Smith, and J. York, February 2003. Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76– 80. ISSN 1089-7801.

David G. Lowe, 1999. Object recognition from local Scale-Invariant features. In *Computer Vision, IEEE International Conference on*, volume 2, page 1150, Los Alamitos, CA, USA. IEEE Computer Society. ISBN 0-7695-0164-8.

B. S. Manjunath, Jens-Rainer Ohm, Vinod V. Vasudevan, and Akio Yamada, 2001. Color and texture descriptors. *IEEE Transactions on circuits and systems for video technology*, 11(6): 703–715. ISSN 1051-8215.

C.D. Manning, Prabhakar Raghaven, and Hinrich Schütze, April 2009. *An Introduction to Information Retrieval.* Cambridge University Press, online edition.

B. M Marlin, R. S Zemel, S. Roweis, and M. Slaney, 2007. Collaborative filtering and the missing at random assumption. In *Proceedings of the 23rd Conference*, volume 47, page 50–54.

Benjamin M. Marlin, 2008. *Missing data problems in machine learning.* PhD thesis, University of Toronto.

Benjamin M Marlin and Richard S Zemel, 2009. Collaborative prediction and ranking with non-random missing data. In *Proceedings of the third ACM conference on Recommender systems*, RecSys '09, page 5–12, New York, NY, USA. ACM. ISBN 978-1-60558-435-5. ACM ID: 1639717.

Cameron Marlow, Mor Naaman, Danah Boyd, and Marc Davis, 2006. HT06, tagging paper, taxonomy, flickr, academic article, ToRead. In *Proceedings of the seventeenth conference on Hypertext andHypermedia*, Odense, Denmark.

Norbert Martínez-Bazan, Victor Muntés-Mulero, Sergio Gómez-Villamor, Jordi Nin, Mario-A. Sánchez-Martínez, and Josep-L. Larriba-Pey, 2007. Dex: high-performance exploration on large graphs for information retrieval. In *Proceedings of the sixteenth ACM*

conference on Conference on information and knowledge management, pages 573–582, Lisbon, Portugal. ACM. ISBN 978-1-59593-803-9.

Adam Mathes, 2004. Folksonomies - cooperative classification and communication through shared metadata. *Computer Mediated Communication*.

Krystian Mikolajczyk and Cordelia Schmid, 2005. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 27(10):1615–1630.

Yashar Moshfeghi, Deepak Agarwal, Benjamin Piwowarski, and Joemon Jose. Movie recommender: Semantically enriched unified relevance model for rating prediction in collaborative filtering. In *Advances in Information Retrieval*, 2009, pages 54–65. Springer-Verlag Berlin Heidelberg.

H. Müller, S. Marchand-Maillet, and T. Pun, 2002. The truth about corel-evaluation in image retrieval. *Image and Video Retrieval*, page 38–49.

Radu Negoescu, July 2007. An analysis of the social network of flickr. *Internal Publication - École Polytechnique Fédérale de Lausanne*.

Radu-Andrei Negoescu, Brett Adams, Dinh Phung, Svetha Venkatesh, and Daniel Gatica-Perez, 2009. Flickr hypergroups. In *Proceedings of the seventeen ACM international conference on Multimedia*, pages 813–816, Beijing, China. ACM. ISBN 978-1-60558-608-3.

Radu Andrei Negoescu and Daniel Gatica-Perez, 2008a. Analyzing flickr groups. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 417–426, Niagara Falls, Canada. ACM. ISBN 978-1-60558-070-8.

Radu Andrei Negoescu and Daniel Gatica-Perez, 2008b. Topickr: flickr groups and users reloaded. In *Proceeding of the 16th ACM international conference on Multimedia*, pages 857–860, Vancouver, British Columbia, Canada. ACM. ISBN 978-1-60558-303-7.

Oded Nov, Mor Naaman, and Chen Ye, 2008. What drives content tagging: the case of photos on flickr. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human*

factors in computing systems, pages 1097–1100, Florence, Italy. ACM. ISBN 978-1-60558-011-1.

The Stationery Office. The british library thirty-seventh annual report and accounts 2009/10. http://www.official-documents.gov.uk/document/hc1011/hc01/0153/0153.asp, July 2010.

D. Olinic, S. Nedevschi, C. Feier, Z. Gal, and N. Olinic, 1999. A structured medical text field of DICOM 3.0 transesophageal echocardiography image file for database implementation. In *Computers in Cardiology 1999*, pages 443–446.

Christopher Olston, Benjamin Reed, Utkarsh Srivastava, Ravi Kumar, and Andrew Tomkins, 2008. Pig latin: a not-so-foreign language for data processing. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1099–1110, Vancouver, Canada. ACM. ISBN 978-1-60558-102-6.

José Luis Ortega and Isidro F. Aguillo, 2008. Análisis estructural de una red social en línea: la red española de flickr. *Profesional de la Información*, 17(6):603–610.

Photobucket. About photobucket. http://photobucket.com/about, October 2010.

Son Lam Phung, Abdesselam Bouzerdoum, and Douglas Chai, 2005. Skin segmentation using color pixel classification: Analysis and comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(1):148–154. ISSN 0162-8828.

John C Platt, 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, pages 61—74.

Michalis Potamias, Francesco Bonchi, Carlos Castillo, and Aristides Gionis, 2009. Fast shortest path distance estimation in large networks. In *Proceeding of the 18th ACM conference on Information and knowledge management*, CIKM '09, page 867–876, Hong Kong, China. ACM. ISBN 978-1-60558-512-3. ACM ID: 1646063.

John Ross Quinlan, 1993. *C4.5: programs for machine learning*. Morgan Kaufmann. ISBN 9781558602380.

Adam Rae, Börkur Sigurbjörnsson, and Roelof van Zwol, April 2010. Improving tag recommendation using social networks. In *9th international conference on Adaptivity, Personalization and Fusion of Heterogeneous Information*.

William J. Reed, December 2001. The Pareto, Zipf and other power laws. *Economics Letters*, 74(1):15–19. ISSN 0165-1765.

S. Rosenthal, M. Veloso, and A. Dey, 2009. Asking questions and developing trust. In *Proceedings of the Spring Symposium on Agents that Learn from Human Teachers*.

G. Salton, A. Wong, and C. S. Yang, 1975. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.

Jose San Pedro and Stefan Siersdorfer, April 2009. Ranking and classifying attractiveness of photos in folksonomies. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 771–780, Madrid, Spain. ISBN 978-1-60558-487-4.

Mark Sanderson and Justin Zobel, 2005. Information retrieval system evaluation: effort, sensitivity, and reliability. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 162–169, Salvador, Brazil. ACM. ISBN 1-59593-034-5.

Patrick Schmitz, 2006. Inducing ontology from flickr tags. In *WWW, May 22–26, Edinburgh, UK*.

Shilad Sen, F. Maxwell Harper, Adam LaPitz, and John Riedl, 2007. The quest for quality tags. In *Proceedings of the 2007 international ACM conference on Supporting group work*, GROUP '07, page 361–370, New York, NY, USA. ACM. ISBN 978-1-59593-845-9. ACM ID: 1316678.

Shilad Sen, Jesse Vig, and John Riedl, 2009. Tagommenders. In *Proceedings of the 18th international conference on World wide web - WWW '09*, page 671, Madrid, Spain.

Zack Sheppard. 5,000,000,000 - flickr blog. http://blog.flickr.net/en/2010/09/19/5000000000/, September 2010.

Börkur Sigurbjörnsson and Roelof van Zwol, 2008. Flickr tag recommendation based on collective knowledge. In *Proceeding of the 17th international conference on World Wide Web*, pages 327–336, Beijing, China. ACM. ISBN 978-1-60558-085-2.

Sanjay Kr. Singh, D.S. Chauhan, Mayank Vatsa, and Richa Singh, 2003. A robust skin colour based face detection algorithm. *Tamkang Journal of Science and Engineering*, 6(4):227–234.

Adish Singla and Ingmar Weber, 2009. Camera brand congruence in the flickr social graph. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining - WSDM '09*, page 252, Barcelona, Spain.

Alan F. Smeaton, Paul Over, and Wessel Kraaij, 2006. Evaluation campaigns and TRECVid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, page 321–330, New York, NY, USA. ACM Press. ISBN 1-59593-495-2.

Karen Sparck Jones, 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Alina Stoica and Christophe Prieur, 2009. Structure of neighborhoods in a large social network. In *Computational Science and Engineering, IEEE International Conference on*, volume 4, pages 26–33, Los Alamitos, CA, USA. IEEE Computer Society. ISBN 978-0-7695-3823-5.

Apostolos Syropoulos, 2001. Mathematics of multisets apostolos syropoulos. *Multiset Processing*, 2235:347—358.

Hideyuki Tamura, Shunji Mori, and Takashi Yamawaki, 1978. Textural features corresponding to visual perception. *Systems, Man and Cybernetics, IEEE Transactions on*, 8(6):460–473. ISSN 0018-9472.

J. Tang and P. H Lewis, 2007. An image based feature space and mapping for linking regions and words. In *Proceedings of 2nd International Conference on Computer Vision Theory and Applications*, page 29–35.

Chayant Tantipathananandh, Tanya Berger-Wolf, and David Kempe, 2007. A framework for community identification in dynamic social networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, San Jose, California, USA.

Jaime Teevan, Meredith Ringel Morris, and Steve Bush, 2009. Discovering and using groups to improve personalized search. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 15–24, Barcelona, Spain. ACM. ISBN 978-1-60558-390-7.

M. Van Erp and L. Schomaker, 2000. Variants of the borda count method for combining ranked classifier hypotheses. In *Proceedings of the Seventh International Workshop on Frontiers in Handwriting Recognition*, page 443–452.

C Van Rijsbergen, 1979. *Information retrieval*. Butterworths, 2d ed. edition. ISBN 9780408709293.

Roelof van Zwol, November 2007. Flickr: Who is looking? *Proceedings of the IEEE/WIC/ACM International Confeence on Web Intelligence*, pages 184–190.

Roelof van Zwol, Adam Rae, and Lluis Garcia Pueyo, 2010. Prediction of favourite photos using social, visual, and textual signals. In *Proceedings of the international conference on Multimedia*, MM '10, pages 1015–1018, New York, NY, USA. ACM. ISBN 978-1-60558-933-6. URL http://doi.acm.org/10.1145/1873951.1874138.

Vladimir Vezhnevets, Vassili Sazonov, and Alla Andreeva, 2003. A survey on Pixel-Based skin colour detection techniques. *Proceedings of Graphicon 2003*.

Lizhe Wang, Gregor Laszewski, Andrew Younge, Xi He, Marcel Kunze, Jie Tao, and Cheng Fu, June 2010. Cloud computing: a perspective study. *New Generation Computing*, 28(2): 137–146. ISSN 0288-3635.

D.J. Watts and S.H. Strogatz, June 1998. Collective dynamics of 'small-world' networks. *Nature*, 393.

M. Williamson, 2009. The latent imager. *Engineering & Technology*, 4(14):36–39. ISSN 1750-9637.

Ian H. Witten and Eibe Frank, 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2nd edition.

Lisa Wiyartanti and Laehyun Kim, 2009. Finding similar music artists for recommendation. In *Proceedings of IADIS International Conference*, volume 1, pages 535–542. ISBN 978-972-8924-93-5.

Filip Jay Yeskel. High volume document image archive system and method, September 2000. US Patent 6,115,509.

Tom Chao Zhou, Hao Ma, Irwin King, and Michael R. Lyu, 2009. TagRec: leveraging tagging wisdom for recommendation. In *Computational Science and Engineering, IEEE International Conference on*, volume 4, pages 194–199, Los Alamitos, CA, USA. IEEE Computer Society. ISBN 978-0-7695-3823-5.

++?????++ Out of Cheese Error. Redo From Start.

(Terry Pratchett, Interesting Times)